Unraveling the Evolutionary History of Orangutans (genus: *Pongo*)

—

The Impact of Environmental Processes and the Genomic Basis of Adaptation

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Maja Patricia Mattle-Greminger

von

Richterswil (ZH)

Promotionskomitee

Prof. Dr. Carel van Schaik (Vorsitz)

PD Dr. Michael Krützen (Leitung der Dissertation)

Dr. Maria Anisimova

Zürich, 2015

To my family

Table of Contents

Summary

Understanding how adaptive variation in phenotypic traits evolved by natural selection and contributed to the process of speciation is one of the central goals of evolutionary biology. Until recently, however, identifying the genetic basis and mechanisms underlying adaptation at the genome-wide scale was technologically beyond reach. The advent of high-throughput sequencing has transformed our ability to address basic questions of evolutionary genetics such as the relative importance of the processes shaping patterns of genetic variation within and among species.

Orangutans (genus: *Pongo*) are the only Asian great apes and currently endemic to the Sundaland islands of Borneo (*P. pygmaeus*) and Sumatra (*P. abelii*). The Southeast Asian Sunda archipelago is a tropical hotspot of biodiversity with unusually high level of species endemism, making it to one of the most exciting global regions in which to investigate how environmental processes lead to local adaptation and species diversification. The region has been drastically affected by geological and environmental processes such as tectonic plate movements, Quaternary climatic oscillations, fluctuating sea levels, and volcanic eruptions. The Sunda Shelf was for instance cyclically exposed during glacial periods when sea levels were lower, which repeatedly reconnected the islands.

Orangutans show remarkable, well-documented geographic variation in various traits related to morphology, physiology, life history, behavioral ecology, and social organization, suggesting high levels of local adaptions. A considerable part of this variation is almost certainly linked to environmental differences throughout the genus' range. The variation in phenotypic traits offers a great opportunity to investigate the interaction between environmental forces and genetic variation both between and within species. In this dissertation, I studied the evolutionary history of the genus *Pongo*, i.e. how environmental processes in Sundaland shaped patterns of genetic variation, and the genetic basis underlying local adaptations, using genome-wide data.

I focused in the first part of this dissertation on methodological aspects, i.e. how to generate suitable population genomic data. The mammalian Y chromosome is critical for studying male-specific evolutionary histories. Yet, due its complex architecture, Y-linked data have remained elusive for most mammals and large-scale data are available mainly for humans. I first reviewed the current and emerging methodological strategies applied to developing Y-specific markers, and subsequently developed a novel bioinformatics strategy, widely applicable to other mammal species, to extract Y-specific single-copy sequences from whole-genome sequencing data. This approach allowed me to comparatively trace both the male- and female-specific evolutionary history on a genomic level for the first time in a non-human great ape. My results demonstrate the great importance and power of genomic Y-specific data for the comprehensive understanding of a species' evolutionary history. Furthermore, I developed a protocol for improved reduced genome complexity sequencing, which allows

sampling only a fraction of the genome with very high genotyping-by-sequencing efficiency and reproducibility among samples. My method is part of a growing suite of similar strategies that have transformed our ability to generate genomic data from natural populations.

In the second and main part of this dissertation, I investigated the evolutionary history of orangutans based on a unique dataset of autosomal and sex-specific whole-genome data covering the genus' entire current geographic range. I found that orangutan evolutionary history is a tale of two islands and two sexes, shaped by the highly dynamic environmental conditions on the Sunda archipelago, and the pronounced sex-biased dispersal in this genus. The speciation of Bornean and Sumatran orangutans was a gradual process over several hundreds of thousands of years, starting in the early Pleistocene, and was heavily influenced by recurrent climate changes and high levels of male-biased dispersal and strict female philopatry. I estimated cessation of gene flow between species to be considerably earlier than proposed previously (~0.43 Ma).

My findings further revealed that the two orangutan species were affected differently by the Pleistocene climate oscillations. Climatic and thus rainforest cover oscillations had a major impact on Bornean orangutans, causing repeated bottlenecks (including a common rainforest refugium in the late Pleistocene) and a long-term population decline. In contrast, Sumatran orangutans were much less affected by climate changes and experienced a remarkably stable population history and structure throughout the Pleistocene. Only recently, they also faced a drastic population decline, most likely caused by the Toba supereruption ~73 ka and prehistoric hunting by early hunter-gatherers. The former adds to the controversial discussion about the consequences of the Toba supereruption by providing, to my knowledge, the first direct evidence of a strong regional impact of the supereruption on a large mammal.

Finally, I present the first genome-wide scans aiming at detecting positive selection within the genus *Pongo*. My results suggest that Bornean orangutans, particularly those in the northeast of the island (*P. p. morio*), may exhibit genetic adaptations to cope with strong fluctuations of fruit abundance and prolonged lean periods in conjunction with unpredictable El Niño events and climate oscillations. For instance, I found signals of potential adaptation pertaining to energy storage (i.e. adipose tissue) metabolism. This is in line with the observed greater ability of Bornean orangutans to deposit large fat storages compared to Sumatran orangutans, which is assumed to allow for physiological buffering against starvation. Furthermore, I identified several candidate genes and biological processes related to neurogenesis, which is consistent with the smaller brain size of the northeastern Bornean orangutans and may again represent an adaptation to survive periods of food scarcity by reducing costs of metabolically expensive brain tissue. In contrast, in Sumatran orangutans, which do not face the same environmental constrains and have more favorable energy budgets, I found signatures of potential adaptive evolution within genes related to learning and adult brain plasticity, the oxytocin pathway, heart development, and hearing. I hypothesize that selective changes in these genes may provide Sumatran orangutans a framework for extended behavioral plasticity linked to their larger and more complex cultural

repertoire and their higher sociability. Overall, my results suggest that at least some of the striking geographic variation in orangutan phenotypic traits may indeed represent genetic local adaptations.

The findings of this dissertation also have important ramifications for the taxonomy and conservation management of orangutans. For instance, in light of the long-lasting separation of orangutans to the south and to the north of Lake Toba in Sumatra, I suggest a taxonomic revision of *P. abelii*. My results further provide the first comprehensive assessment of conservation units within *Pongo*.

Zusammenfassung

Ein grundlegendes Ziel der Evolutionsbiologie ist zu verstehen, wie adaptive Variation phänotypischer Merkmale infolge natürlicher Selektion entstanden ist und zur Artbildung beigetragen hat. Bis vor kurzem war es jedoch technologisch unerreichbar, die genetische Basis und die Mechanismen, die dem Adaptationsprozess zugrunde liegen, in einem genomweiten Ansatz zu untersuchen. Das Aufkommen des sogenannten "Next Generation Sequencing" hat unsere Möglichkeiten fundamentale Fragestellungen in der Evolutionären Genetik zu untersuchen revolutioniert, wie beispielsweise die Frage nach dem relativen Einfluss verschiedener evolutiver Prozesse auf die Musterbildung genetischer Variation innerhalb und zwischen Arten.

Orang-Utans (Gattung: *Pongo*) sind die einzigen asiatischen Menschenaffen und kommen heutzutage nur noch in den Regenwäldern auf den Sundainseln Borneo (*P. pygmaeus*) und Sumatra (*P. abelii*) vor. Die südostasiatische Inselgruppe Sundaland ist ein tropischer Hotspot an Biodiversität mit einem aussergewöhnlich hohen Anteil endemischer Arten. Dies macht Sundaland weltweit zu einer der spannendsten Regionen, um zu untersuchen, wie Umweltprozesse zu lokaler Adaption und zur Diversifikation von Arten führen. Während des Quartärs stand die Region unter ausgesprochen starkem Einfluss klimatischer Schwankungen. Das Kontinentalschelf war beispielsweise infolge des sinkenden Meeresspiegels während Eiszeiten periodisch freiliegend, was wiederholt zu Landbrücken zwischen den Inseln führte.

Orang-Utans zeigen aussergewöhnliche und gut dokumentierte geografische Variation in vielerlei Merkmalen betreffend Morphologie, Physiologie, Lebenszyklus, Verhaltensökologie und Sozialorganisation, was auf ein hohes Mass an lokaler Adaption hindeutet. Ein beträchtlicher Anteil dieser Variation ist mit grosser Wahrscheinlichkeit gekoppelt an die ökologischen Unterschiede innerhalb des Verbreitungsgebiets der Gattung *Pongo*. Die Variation in phänotypischen Merkmalen bietet eine ausgezeichnete Möglichkeit, die Einwirkung von Umweltprozessen auf die genetische Variation zwischen und innerhalb von Arten zu studieren. In dieser Dissertation untersuchte ich die Evolutionsgeschichte der Orang-Utans und die genetische Basis, die lokalen Adaptionen zugrunde liegt, mittels einem das ganze Genom umfassendem Ansatz.

Im ersten Teil dieser Dissertation legte ich den Fokus auf methodische Aspekte, i. e. wie adäquate populations-genomische Daten generiert werden können. Das Y-Chromosom ist bei Säugetieren von entscheidender Bedeutung um die Männchen-spezifische Evolutionsgeschichte zu untersuchen. Aufgrund der hochkomplexen Struktur des Y-Chromosoms ist es jedoch äussert schwierig genetische Marker zu entwickeln und grössere Datenmengen sind deshalb bis jetzt praktisch nur vom Mensch vorhanden. Im Kontext eines Reviews habe ich als erstes die gängigen methodologischen Strategien zur Entwicklung von Y-spezifischen Markern geprüft und neue mögliche Strategien erläutert. Daraufhin habe ich eine neuartige, breit auf Säugetiere anwendbare, bioinformatische Strategie ausgearbeitet,

die es ermöglicht, Y-spezifische Sequenzen von Ganz-Genom-Sequenzierungsdaten zu gewinnen. Mit dieser Methode war es mir möglich, erstmals die Männchen- und Weibchen-spezifische Evolutionsgeschichte eines Menschenaffen – abgesehen vom Menschen selbst – auf genomischer Ebene zu vergleichen und nachzuvollziehen. Meine Resultate zeigen die grosse Bedeutung und das Potenzial von genomischen Y-spezifischen Daten für das umfassende Verständnis der Evolutionsgeschichte einer Art. Des Weiteren entwickelte ich ein Protokoll für ein verbessertes reduziertes-Genomkomplexitäts-Sequenzieren, das erlaubt kostengünstig nur einen Teil des Genoms zu sequenzieren und dies mit einer sehr hohen Genotypisierung-durch-Sequenzierung Effizienz und Reproduzierbarkeit zwischen genetischen Proben. Mein Protokoll ist Teil eines wachsenden Methodenpaktes, das uns in die Lage versetzt, Genomdaten von natürlichen Populationen zu generieren.

Im Hauptteil dieser Dissertation untersuchte ich die Evolutionsgeschichte der Orang-Utans und wie diese durch die dynamischen Umweltprozesse in Sundaland beeinflusst wurde. Dafür erarbeitete ich einen einzigartigen Datensatz autosomaler und geschlechts-spezifischer Genomdaten, der das gesamte derzeitige geografische Verbreitungsgebiet der Gattung *Pongo* abdeckt. Ich stellte fest, dass die Evolutionsgeschichte der Orang-Utans eine Geschichte von zwei Inseln und zwei Geschlechter ist, geprägt durch stark schwankende Umweltbedingungen in Sundaland und aussergewöhnlich unterschiedlichen geschlechts-spezifischen Abwanderungstendenzen von Männchen und Weibchen in dieser Gattung. Die Artbildung von *P. abelii* und *P. pygmaeus* war ein gradueller Prozess über mehrere hunderttausende von Jahren beginnend anfangs des Pleistozäns. Sie stand unter starkem Einfluss wiederkehrender klimatischer Veränderungen und einem hohen Ausmass an Männchen-spezifischem Genfluss sowie strikter Philopatrie der Weibchen. Meine Resultate deuten darauf hin, dass der Genfluss zwischen den beiden Arten deutlich früher als bisher angenommen geendet hat (vor ca. 430'000 Jahren).

Des Weiteren offenbarten meine Forschungsergebnisse, dass die klimatischen Schwankungen während des Pleistozäns die beiden Orang-Utan Arten unterschiedlich beeinflusst haben. Klimaveränderungen und die damit verbundenen Schwankungen des Erstreckungsgebietes des Regenwaldes hatten einen wesentlichen Einfluss auf die Orang-Utans auf Borneo. Die Fluktuationen führten zu wiederholten genetischen Flaschenhals-Effekten (mitunter zu einem gemeinsamen glazialen Refugium während des späten Pleistozäns) und einem seit langem andauernden Populationsrückgang. Im Gegensatz dazu waren die Orang-Utans auf Sumatra weit weniger von den klimatischen Schwankungen betroffen und durchlebten eine auffallend stabile Populationsgeschichte und Populationsstruktur während des gesamten Pleistozäns. Erst in jüngerer Vergangenheit erfuhren sie ebenfalls einen drastischen Populationsrückgang, der sehr wahrscheinlich auf die Supereruption des Toba-Vulkans vor ca. 73'000 Jahren und auf die Bejagung durch prähistorische Jäger und Sammler zurückzuführen ist. Der – meines Wissens erstmalige – unmittelbare Beleg eines gewichtigen regionalen Einflusses der Supereruption auf einen grossen Säuger trägt wesentlich zur kontroversen Diskussion über die Folgen der Toba Supereruption bei.

Abschliessend präsentiere ich erstmals genomweite Analysen mit dem Ziel positive Selektion innerhalb der Gattung *Pongo* zu detektieren. Meine Resultate deuten darauf hin, dass Orang-Utans auf Borneo, insbesondere diejenigen im Nordosten der Insel (*P. p. morio*), genetische Adaptionen aufweisen, um starke Fluktuationen in der Nahrungsmenge mit längeren Zeiträumen von ausgeprägter Nahrungsknappheit, zu bewältigen. Diese treten vor allem im Zusammenhang mit El Niño-Perioden auf und betreffen den Nordosten von Borneo am stärksten. Ich habe beispielsweise genetische Signale gefunden, die auf mögliche Adaptation betreffend des Energiespeicherungs-Metabolismus (im Spezifischen des Stoffwechsels der adipösen Zellen) hindeuten. Dies steht im Einklang mit der Beobachtung, dass die Orang-Utans Borneos bei Nahrungsüberfluss weit grössere Fettreserven anlegen als die Orang-Utans Sumatras. Es wird angenommen, dass dieses dem physiologischen Puffern in Hungerperioden dient. Ich habe zudem mehrere Kandidaten-Gene und biologische Prozesse identifiziert die mit der Neurogenese zusammenhängen, was kongruent zu der kleineren Gehirngrösse der Orang-Utans im Nordosten von Borneo ist. Diese stellt möglicherweise ebenfalls eine Adaption dar, womit das Überleben bei extremer Nahrungsknappheit gesichert wird, indem der energetische Aufwand des metabolisch anspruchsvollen Hirngewebes gesenkt wird.

Orang-Utans auf Sumatra sind dagegen deutlich weniger harschen Umweltbedingungen ausgesetzt. Im Einklang damit habe ich dementsprechend andere Signale von möglicher positiver Selektion gefunden. Diese betrafen zum Beispiel Gene, die wichtige Funktionen haben in Bezug auf die Lernfähigkeit und die Plastizität des adulten Gehirns, den Oxytocin Signalweg, die Herzentwicklung und das Gehör. Ich stelle die Hypothese auf, dass selektive Änderungen in diesen Genen den Orang-Utans auf Sumatra die Rahmenbedingungen zur Erweiterung plastischer Verhaltensweisen – verbunden mit ihrem weit grösseren und komplexeren kulturellen Repertoire und ihrer höheren Sozialität – bieten. Insgesamt deuten meine Resultate darauf hin, dass zumindest ein Teil der bemerkenswerten geographischen Variation in phänotypischen Merkmalen in Orang-Utans genetische lokale Adaptionen darstellen.

Die Erkenntnisse dieser Dissertation haben wesentliche Auswirkungen für die Taxonomie und den Artenschutz von Orang-Utans. Ich schlage unteranderem vor, die Taxonomie von *P. abelii* in Anbetracht der langandauernden Separation von Orang-Utans nördlich und südlich vom Tobasee in Sumatra zu revidieren. Meine Resultate erlauben zudem die erste umfassende Beurteilung von sogenannten "conservation units" innerhalb der Gattung *Pongo*.

Acknowledgments

I began this dissertation because I was fascinated by the new avenues next generation sequencing has opened up to the field of evolutionary biology by finally making answers accessible to long-standing questions about adaptive evolution. It has been extremely exciting, and also challenging, to witness the incredibly rapid development of the young field of genomics. A large number of people and institutions have supported me and my research endeavors in various ways for which I am very grateful.

First and foremost, I thank my advisor and mentor Michael Krützen. At the time I had begun my PhD, the field of genomics was still its infancy years and we did not know to what extend this project would succeed. Notwithstanding, Michael enabled me to carrying out this dissertation. He always greatly supported me, believed in me and encouraged me no matter which challenges I faced. I am particularly grateful for his continuous, invaluable support during the time of my health issues. Thank you Michael for all you did for me and the fantastic time we had together.

I am also very grateful to Carel van Schaik for his valuable inputs and support, and for being a constant source of inspiration and enthusiasm. His knowledge in many different fields of science has never stopped amazing me. The in-depth knowledge about orangutan behavioral ecology Carel, Maria van Noordwijk, and numerous other fellow orangutan researchers gathered over many years of field research ultimately provided the basis for studying orangutan adaptive evolution and I am indebted to all of them.

Special thanks go to Maria Anisimova for her enthusiasm about my work, the great discussions, the valuable advices, and the enjoyable moments together. Thank you for being a member of my committee.

Furthermore, I would like to thank Tomas Marques-Bonet for our highly fruitful and pleasant collaboration. It is very inspiring working with you. I highly value our discussions and your inputs. I am also grateful to Tomas and Javier Prado-Martinez for sharing their unpublished genome sequencing data and for their efforts related to the sequencing of the novel individuals.

Many thanks go to Peter Wandeler for being my great companion in the quest for the Y and for the important contributions as a senior author in our joint publication. I highly appreciated and enjoyed our collaboration.

I also highly valued the collaboration with Kai Stölting. Our intensive exchange about RAD/RRL sequencing and bioinformatics has been a great help. Many thanks for your support and the nice days in Fribourg.

Furthermore, I am deeply grateful to Alexander Nater, who had a major impact on especially the last three data chapters presented in this dissertation. Many thanks for the countless discussions and long Skype sessions, your invaluable feedback, the legendary perl pipelines, sharing your great theoretical knowledge, commenting on all chapters of this dissertation, and for your friendship.

I am indebted to Benoit Goossens, Reeta Sharma, Laurentius Ambu, Lounes Chikhi, Sen Natalan, the Sabah Wildlife Department, and the staff at the Sepilok Orangutan Rehabilitation Centre, the Shangri-La's Rasa Ria Resort, and the Lok Kawi Wildlife Park, for their help with collecting orangutan samples in Sabah in 2010. I am particularly grateful to Benoit Goossens who made this sampling possible and provided invaluable support. The samples that we could collect proved to be crucial for the success of this project. Very special thanks go to Reeta Sharma, my field companion, for sharing all ups and downs, and for the unforgettable time we spend together in Sabah. I am also especially thankful to the veterinarian Cecilia Brooklin for her help as well as to the former director of the Sabah Wildlife Department Laurentius Ambu for his generous hospitality and support.

Furthermore, I am grateful to Ernst Verschoor and Kristin Warren for contributing important orangutan samples from Kalimantan. I would also like to thank Ian Singleton, Joko Pamungkas, Dyah Perwitasari-Farajallah, Muhammad Agil, and the staff at the Sumatran Orangutan Conservation Programme, BOS Wanariset Orangutan Reintroduction Project, and Semongok Wildlife Rehabilitation Centre for their efforts related to sampling and/or obtaining permits. I also thank the following institutions for supporting this research: Sabah Wildlife Department (SWD), Indonesian State Ministry for Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), Leuser International Foundation (LIF), Taman National Gunung Leuser (TNGL), and Borneo Orangutan Survival Foundation (BOSF). In addition, I would like to thank Werner Schempp and his team for the great efforts they invested trying to establish cell lines from orangutan blood samples.

I further would like to thank Rémy Bruggmann, Andrea Patrignani, and the staff at the Functional Genomics Centre Zurich for their support related to sequencing. I also much appreciated discussions with Beatrice Nussberger and Robert Kraus about reduced-representation sequencing methods. In addition, I thank Ivo Gut and Marta Gut, and the team at the Centro Nacional de Análisis Genómico for carrying out whole-genome sequencing.

I am particularly grateful to David Marques for the numerous discussions about how studying adaptive evolution and for general support especially at the beginning of my PhD. Very special thanks go to Giada Ferrari for the work she carried out during her internship. I also would like to thank Heidi Lischer for sharing scripts and for discussions. Furthermore, I am grateful to Christian Roos for contributing unpublished mitogenome sequences. I also thank the S3IT team for the extended access to the Schroedinger HPC cluster.

I highly appreciated discussions with Christian Lexer, Gerald Heckel, Mike Bruford, Laurent Excoffier, Aylwyn Scally, Heng Li, Johannes Krause, George (PJ) Perry, Lounès Chikhi, Lukas Keller, Pablo Orozco-terWengel, and many others.

Many warm thanks to Claudia Zebib for her administrative and personal support throughout all these years, Ruth Haegi for help with all sorts of administrative problems, as well as Marcus Gisi for help with IT issues. I also thank Tony Weingrill, who runs the PhD Program in Evolutionary Biology.

Very special thanks go to the members of the Evolutionary Genetics Group (EGG). I always felt very fortunate to be working in such a team-oriented group with such a fantastic spirit (and baking abilities). Thank you all for the great time! I want to especially thank my (former) office mates Alexander Nater, Natasha Arora, Anna Kopps, Livia Gerber, Kathrin Bacher, Nadja Morf, David Marques, Pirmin Nietlisbach, Maria Jendensjö, and Corinne Ackermann.

I would also like to thank everyone else at the Anthropological Institute and Museum. I highly appreciated and enjoyed the great and stimulating working atmosphere, the countess (scientific and non-scientific) discussions, encouragement, and friendship.

I would gratefully like to acknowledge major financial support from the A.H. Schultz Stiftung (to MK and MPG), the Forschungskredit by the University of Zurich (to MPG), the Leakey Foundation (to MPG), the UZH University Research Priority Program (to MK), Julius-Klaus Foundation (to MK), and Swiss National Science Foundation (grant no. 3100A-116848 to MK and CPvS). I would particularly like to express my gratitude to Wolfgang Zenker und Hans-Konrad Schmutz for greatly supporting this work and myself.

Finally, this dissertation would not have been possible without the endless support of my beloved family and friends. There are no words great enough to thank all of you.

Chapter 1

General introduction

1.1 Evolutionary processes shaping patterns of genetic diversity

One of the ultimate goals of evolutionary biology is to understand how adaptive variation in phenotypic traits evolved by natural selection. Although Darwin and Wallace (1858) already postulated in the nineteenth century that evolutionary change is governed by the spread and subsequent fixation of favorable traits, empirical demonstrations of selection in natural populations have proven difficult (for a review see Savolainen *et al.* 2013). Identifying the genetic basis and mechanisms underlying local adaptations has been a long-standing endeavor (Fisher 1930; Kimura 1984; Orr 1998). Main questions include: what is the genetic architecture of adaptive change? Is adaptation more likely to occur from new mutations arising or from standing genetic variation? To what extend is the genetic basis shared in case of convergent evolution? What is the relative influence of adaptive and purifying selection in molecular evolution?

Current patterns of genetic variation within and among species are the result of demographic, selective and stochastic processes during the course of a species' evolutionary history (Wall *et al.* 2002; Haddrill *et al.* 2005; Nielsen *et al.* 2005b; Stajich & Hahn 2005; Hahn 2008). Unraveling how these different evolutionary processes have shaped the genetic makeup of a species is a main interest of evolutionary genetics (Lewontin 1974; Mayr 1982), and disentangling their relative contributions is challenging (e.g. Nei *et al.* 2010). For instance, certain demographic events, such as population structuring or population size changes, can produce similar signatures in the genome as selection, although they are only due to random genetic drift (Tajima 1989; Andolfatto & Przeworski 2000; Nielsen 2005; Teshima *et al.* 2006; Excoffier *et al.* 2009). Therefore, detailed information about the demographic history and extant population structure is required to disentangle these potentially confounding effects from true signals of selection (Wall *et al.* 2002; Haddrill *et al.* 2005; Nielsen *et al.* 2005b; Stajich & Hahn 2005). Overall, for a comprehensive understanding of a species' evolutionary history and the process of speciation we need knowledge of (i) geographic structuring of gene pools, (ii) demographic history of populations, (iii) genetic adaptions to local habitat conditions, and (iv) the environmental and biological factors shaping all of the aforementioned.

1.2 Processes shaping patterns of genetic variation in orangutans

This dissertation investigates the processes that have shaped patterns of genetic variation in the only Asian great ape. The two currently recognized species of the genus *Pongo* are endemic to the islands of Borneo (*P. pygmaeus*) and Sumatra (*P. abelii*), which both are part of the Sunda archipelago in Southeast Asia. Orangutans show remarkable geographic variation in various traits related to morphology, physiology, life history, behavioral ecology, and social organization (van Schaik *et al.* 2009b; Wich *et al.* 2009b), suggesting high levels of

local adaptions. A considerable part of this variation is almost certainly linked to environmental differences throughout the genus' range. Therefore, the documented variation in phenotypic traits offers a great opportunity to study the interaction between environmental forces and genetic variation both between and within two closely-related great ape species. Due to the basal position of the genus *Pongo* in the lineage leading to African great apes and modern humans (Groves 2001), the genus is of high importance for our understanding of the evolutionary history of great apes in general.

Orangutans have likely experienced a very complex demographic history. The flora and fauna of Sundaland has been drastically impacted by highly dynamic environmental processes during the Quaternary (Hall 2002; Bird *et al.* 2005). Because orangutans are dependent on evergreen rainforest and exhibit an exceptionally slow life history (Delgado & van Schaik 2000), this genus has been suggested to be a model system to study the consequences of environmental processes for species distribution and genetic diversification in Sundaland (e.g. Kanthaswamy & Smith 2002; Steiper 2006; Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2013; Nater *et al.* 2015).

Dynamic environmental processes of Sundaland

Sundaland is a tropical Asian hotspot of biodiversity with unusually high level of species endemism (Myers *et al.* 2000; Sodhi *et al.* 2004), making it to one of the most exciting global regions in which to investigate how geological and environmental processes lead to local adaptation and species diversification. The currently partly submerged shallow Sunda Shelf includes the Malaysian peninsula, the islands of Sumatra, Borneo and Java, as well as several smaller islands (Molengraaff 1921; Bird *et al.* 2005). The Sunda archipelago underwent notable tectonic plate movements causing landmass reconfigurations. The present shape was attained in the Early Pleistocene ~2.6–1.8 million years ago (Ma) (Meijaard 2004). Since then, the region has been severely affected by the Quaternary climatic oscillations (Figure 1; Flenley 1998; Morley 2000; Bird *et al.* 2005; Cannon *et al.* 2009; de Bruyn *et al.* 2014).

The Sunda Shelf was cyclically exposed during glacial periods when sea levels were lower, which repeatedly reconnected the islands (Verstappen 1997; Voris 2000) and potentially allowed for terrestrial migration. Yet, large paleo-river systems dissected the exposed shelf (Figure 2; Rijksen & Meijaard 1999; Harrison *et al.* 2006) and a savanna corridor may have been present, at least around glacial maxima (Bird *et al.* 2005). Both factors have probably imposed substantial barriers to migration of forest-dwelling species between Sundaland islands. Generally, glacial periods were characterized by a considerably more arid and seasonal climate compared to inter-glacials (Morley 2000), leading to strong fluctuations in coverage and elevational distribution of rainforests (Flenley 1998; Morley 2000; Bird *et al.* 2005; Cannon *et al.* 2009; de Bruyn *et al.* 2014). These recurrent habitat expansions and contractions have likely repeatedly isolated and reconnected populations of forest-dwelling species (Gathorne-Hardy *et al.* 2002).

Sundaland was also subjected to extensive volcanic activity, mainly on Sumatra and Java (Hall 1996), which may have led to local extinctions and subsequent recolonization events (Muir *et al.* 2000). Most notable is Mount Toba in northern Sumatra, which had at least four major and numerous smaller eruptions during the Pleistocene (Chesner *et al.* 1991; Hall 1996). The Toba supereruption ~73 thousand years ago (ka) has been the largest volcanic eruption of the Quaternary (Chesner *et al.* 1991). However, the impact of this supereruption on regional and global wildlife remains highly controversial (e.g. Schulz *et al.* 2002; Gathorne-Hardy & Harcourt-Smith 2003; Petraglia *et al.* 2007; Haslam & Petraglia 2010; Williams *et al.* 2010; Williams 2012). Some researches argued that the impact on fauna and flora has actually been limited (e.g. Schulz *et al.* 2002; Gathorne-Hardy & Harcourt-Smith 2003; Petraglia *et al.* 2007; Haslam & Petraglia 2010), while others hypothesized that the eruption induced a 'volcanic winter' that, among others, may have caused a severe population bottleneck in early humans (e.g. Ambrose 1998; Rampino & Ambrose 2000). Finally, the Sundaland region contains a high density of topographical features, such as rivers, lakes, and mountain chains (Rijksen & Meijaard 1999; Voris 2000; Harrison *et al.* 2006), that may have acted as barriers to dispersal and thus have led to geographic structuring of gene pools.

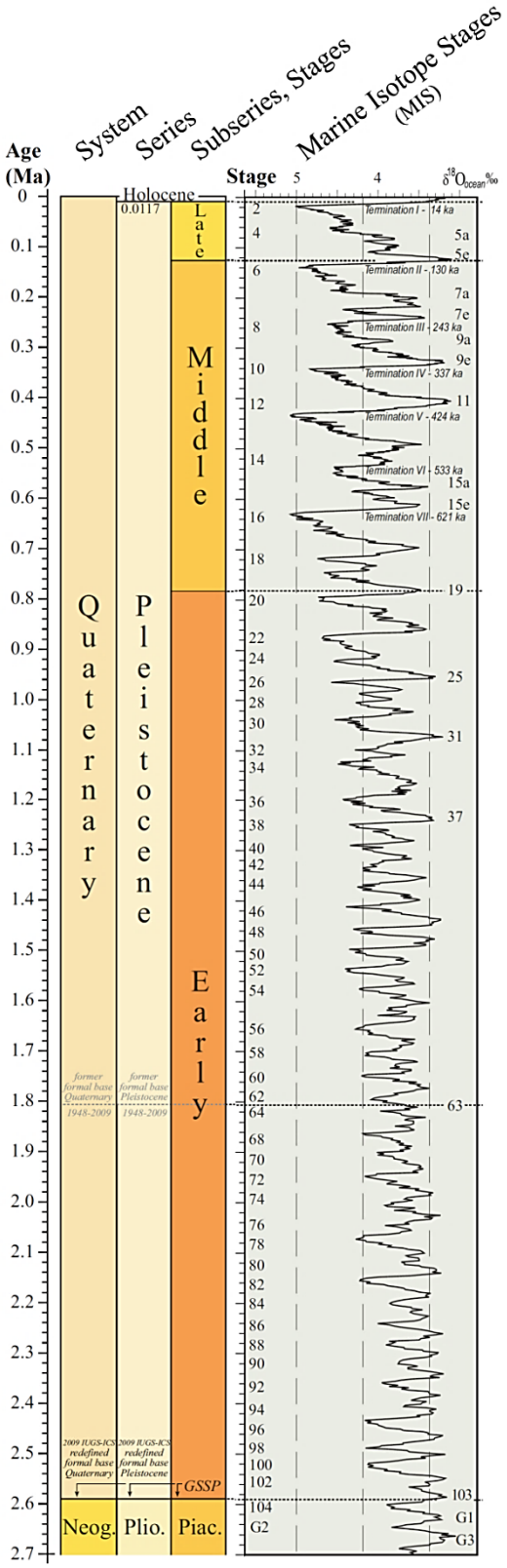


Figure 1. Global climate changes over the past 2.7 million years. Glacial cycles are derived from oxygen isotope data which reflect changes in temperature. Even marine isotope stages (MIS) numbers indicate cold glacial periods, odd MIS numbers denote warm interglacial intervals. Modified from Cohen & Gibbard (2011).

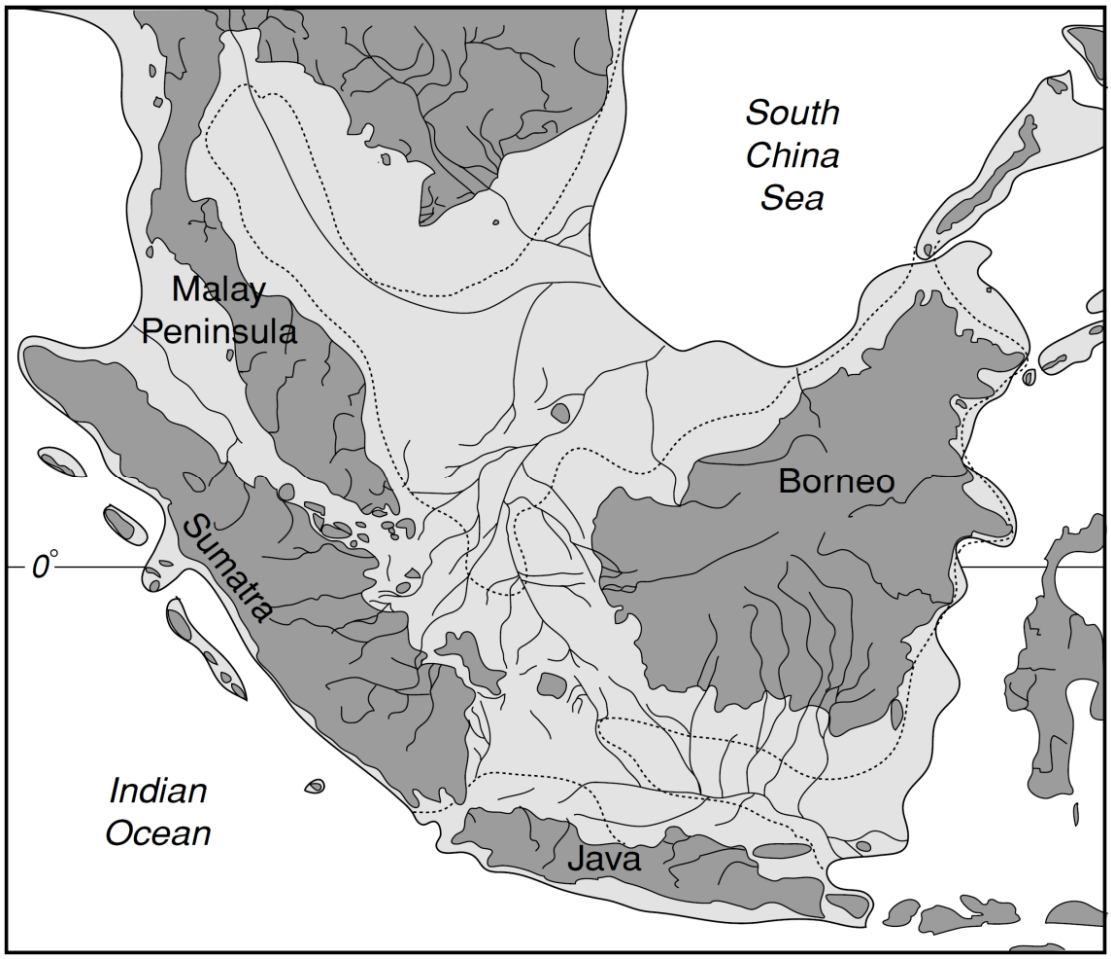


Figure 2. The paleo-river systems on the exposed Sunda Shelf. Solid lines show rivers at sea level 100 meters below present-day, dotted lines at 50 meters below present-day. Taken from Harrison *et al.* (2006).

Distribution and population history of orangutans

The evolutionary history of orangutans has likely been strongly influenced by the aforementioned environmental processes as indicated by major changes in their distribution during the Quaternary (von Koenigswald 1982; Rijksen & Meijaard 1999; Delgado & van Schaik 2000). Fossil records provide evidence that they were once widely distributed throughout mainland Southeast Asia and most of the Sundaland islands (von Koenigswald 1982; Rijksen & Meijaard 1999; Delgado & van Schaik 2000). By the end of the Pleistocene, however, all orangutan populations on the mainland had become extinct (Rijksen & Meijaard 1999; Ibrahim *et al.* 2013). Climatic changes may have led to the southward shift of their distribution (Jablonski 1998; Ibrahim *et al.* 2013). During the Holocene, also the orangutans in southern Sumatra and Java disappeared (Delgado & van Schaik 2000; Ibrahim *et al.* 2013). This has mainly been attributed to anthropogenic factors, i.e. hunting by prehistoric hunter-gatherer societies (Delgado & van Schaik 2000). Nowadays, their range is restricted to increasingly isolated forest patches in northern Sumatra (*P. abelii*), and a wider distribution

on Borneo (*P. pygmaeus*) (Figure 3; Wich *et al.* 2008). Based on morphological characters (Groves 2001) and early genetic data (Warren *et al.* 2001), three subspecies of Bornean orangutans are currently recognized: *P. P. pygmaeus* in northwest Borneo, *P. p. wurmbii* in central-southwest Borneo, and *P. p. morio* in northeast Borneo (Figure 3; Groves 2001; Brandon-Jones *et al.* 2004). No subspecies have been described for Sumatran orangutans.

As all extant non-human great apes, orangutans are highly threatened with extinction, particularly by ongoing habitat loss and fragmentation, as well as illegal hunting and pet trade (Delgado & van Schaik 2000; Goossens *et al.* 2006a; Gaveau *et al.* 2009; Meijaard *et al.* 2011; Wich *et al.* 2012). Sumatran orangutans are listed as critically endangered and Bornean orangutans as endangered (IUCN 2014), with only an estimated 6,600 Sumatran and 54,000 Bornean orangutans left in the wild (Wich *et al.* 2008). Alone within the last century, the census size of orangutans decreased by a least a factor of ten (Rijksen & Meijaard 1999), mostly attributable to industrial deforestation of suitable rainforest habitat.

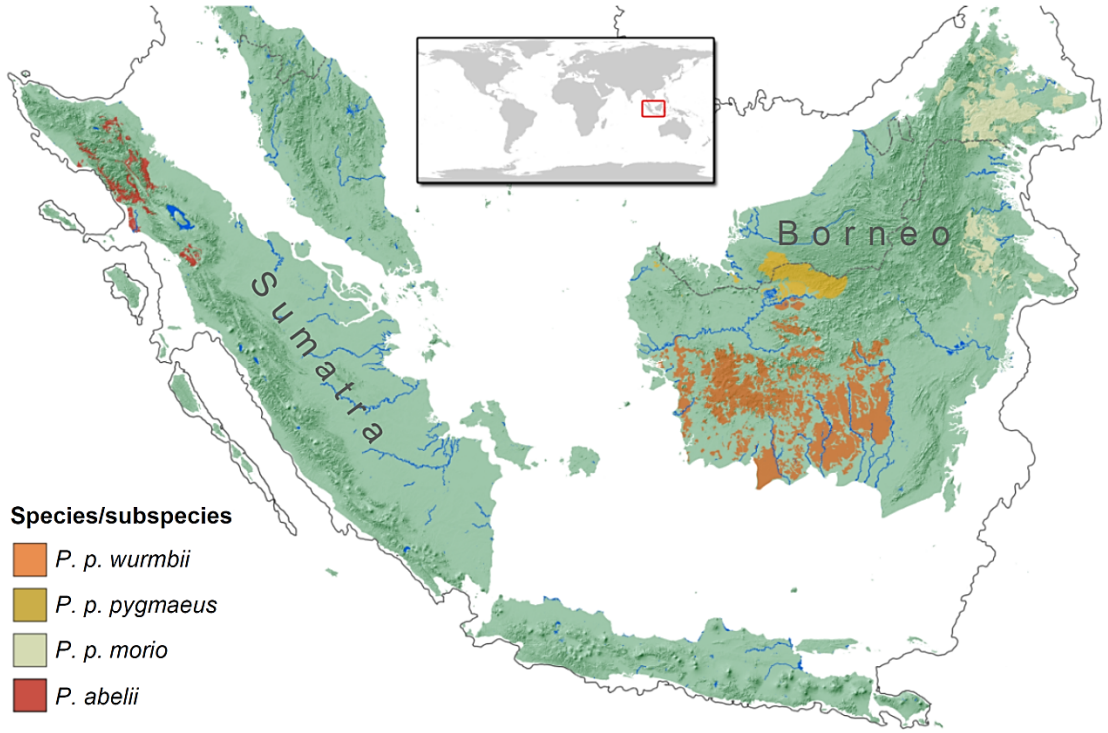


Figure 3. Extant geographic distribution of the genus *Pongo* in Sundaland. The grey line around the islands indicates the extent of the exposed continental shelf during the last glacial maximum (~120 meters below current sea level). Adapted from E.P. Willems.

Signals of the environmental events of Sundaland were found in genetic studies of extant orangutans, pointing towards at a very complex demographic history of the genus *Pongo*. Most studies agree that Sumatran orangutans exhibit much higher genetic diversity and corresponding long-term effective population size (Ne) for autosomal and mitochondrial DNA (mtDNA) than Bornean orangutans (Muir *et al.* 2000; Zhang *et al.* 2001; Steiper 2006; Locke *et al.* 2011; Nater *et al.* 2011; Prado-Martinez *et al.* 2013), despite their much smaller current census size (Wich *et al.* 2008). It had been hypothesized that the larger N_e of Sumatran orangutans was a result of immigration events from different geographic regions (Muir *et al.* 2000; Steiper 2006). However, this scenario has been challenged given the very deep geographic structure of Sumatran mtDNA lineages, which hints at long last-lasting isolation mechanisms preventing admixture of mtDNA lineages (Nater *et al.* 2011). In Bornean orangutans, mtDNA data indicated that their lower genetic diversity may have been related to a severe bottleneck in the late Pleistocene, during which they had been potentially forced into a common rainforest refugium (Arora *et al.* 2010; Nater *et al.* 2011).

In both orangutan species, distinct population structure was documented based on autosomal microsatellite markers and mtDNA loci (Warren *et al.* 2001; Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2013; Nater *et al.* 2015). The particularly complex genetic structure in Sumatra was highlighted by the lack of reciprocal monophyly for mtDNA between both currently recognized orangutan species, where the lineage of Batang Toru—the only remaining Sumatran population south of the Toba caldera—was more closely related to the lineage leading to extant Bornean orangutans than to other Sumatran orangutans (Nater *et al.* 2011).

The divergence time of Bornean and Sumatran orangutans and to what extent the recurrent land bridges between islands facilitated migration remain highly controversial. Studies based on mtDNA loci estimated a divergence time of 1–5 Ma (Zhi *et al.* 1996; Warren *et al.* 2001; Zhang *et al.* 2001; Steiper 2006; Nater *et al.* 2011). Inferences from the autosomal genome indicate a more recent species split time of ~330–600 ka (Locke *et al.* 2011; Mailund *et al.* 2011; Mailund *et al.* 2012). Using Y-linked markers, Nater *et al.* (2011) found an unexpectedly recent coalescence for Bornean and Sumatran orangutans of only 168 ka, suggesting recent gene flow between islands.

The general disagreement between genetic marker systems can partly be attributed to the strongly sex-biased dispersal system of orangutans, where gene flow is almost exclusively male-mediated. Both field and genetic studies have shown that young females preferentially establish their home range overlapping with those of their maternal relatives (Singleton & van Schaik 2002; Arora *et al.* 2012; Nietlisbach *et al.* 2012; van Noordwijk *et al.* 2012). Due to this strong female philopatry, mtDNA will hardly be exchanged among geographic regions. In contrast, males leave their area of birth and intense male-male competition (Utami-Atmoko *et al.* 2009) may force them to travel large distances before being able to establish a new home range (Delgado & van Schaik 2000; Singleton & van Schaik 2001; Nietlisbach *et al.* 2012).

Geographic variation in orangutan behavioral ecology

In a large long-term collaborative effort of orangutan researchers, involving almost all active field sites and hundreds of thousands of hours of observations, a unique data set of behavioral and environmental variation has been compiled for the genus *Pongo* (van Schaik *et al.* 2009b). A comparative synthesis of these findings has recently been published in Wich *et al.* (2009b), revealing a remarkable variation in phenotypic traits. The documented variation (Table 1) largely follows a west–east gradient across the entire range of orangutans from northern Sumatra (*P. abelii*) via western and central Borneo (*P. p. wurmbii*) to eastern and northern Borneo (*P. p. morio*) (van Schaik *et al.* 2009b; Wich *et al.* 2009b; there is insufficient data for *P. P. pygmaeus*). Most of these differences are expected to represent adaptations to ecological variation following the same west–east gradient (Krützen *et al.* 2011; Wich *et al.* 2011b), particularly to habitat productivity and stability of food supply (van Schaik *et al.* 2009b).

Generally, Sumatran rainforests are better habitat for orangutans than Bornean forests (Marshall *et al.* 2009). Forests on northern Sumatra show for instance higher productivity (i.e. mean fruiting rates) than those on Borneo (Husson *et al.* 2009; Marshall *et al.* 2009; Wich *et al.* 2011b). In addition, Sumatran orangutans are normally not exposed to prolonged periods of food scarcity, as fruit availability is temporally more stable in northern Sumatra. In stark contrast, Bornean orangutans have to cope with strong fluctuations in fruit abundance (Wich *et al.* 2006; Morrogh-Bernard *et al.* 2009; Kanamori *et al.* 2010; Wich *et al.* 2011b), also associated with impacts of the El Niño-Southern Oscillation phenomenon (ENSO) (Philander 1983). Particularly in the northeast of Borneo (*P. p. morio*), orangutans are severely affected by the unpredictable ENSO periods (MacKinnon *et al.* 1996; Knott 1998; Delgado & van Schaik 2000). At the intervals of 2–10 years, ENSO events cause prolonged droughts and forest fires that lead to periods of extreme food scarcity. During these periods, orangutans mainly feed on low-energy (fallback) foods such as inner bark, leaves, and other vegetation (Knott 1998; Morrogh-Bernard *et al.* 2009).

In line with the differences in habitat quality, orangutan population densities decrease from west to east (Husson *et al.* 2009; Marshall *et al.* 2009). In the same direction, mandibular robusticity and probably tooth enamel thickness increase, which has been linked to a mechanically more challenging diet with a higher proportion of stiff foods like seeds or inner bark (Taylor 2006; Taylor 2009). A significant decrease in the northeastern *P. p. morio* is observed in both absolute and relative brain size of female orangutans (Taylor & van Schaik 2007; C. P. van Schaik 2010, unpublished data). The smaller brains may represent an adaptation to survive prolonged lean periods (Taylor & van Schaik 2007; van Woerden *et al.* 2012) by reducing costs of metabolically expensive brain tissue ("Expensive Brain framework", Isler & van Schaik 2009; van Woerden *et al.* 2012). The variation in brain size is particularly striking, as it represents, to the best of my knowledge, the only published case of significant differences in brain size found within any great ape species. Although encephalization is seen as a hallmark of the hominid lineage, the genetic basis underlying brain size evolution

remains largely unknown (reviewed in Vallender *et al.* 2008; Enard 2014; Taverna *et al.* 2014; but see Boyd *et al.* 2015).

Orangutans also display systematic physiological differences. Measures of ketone bodies excreted in urine of wild individuals indicate a greater tendency of Bornean orangutans to deposit large fat storages compared to Sumatran orangutans (Knott 1998; Wich *et al.* 2006). This finding is supported by anecdotal evidence from captive orangutans where obesity in response to long-term food abundance is much more commonly seen among Bornean than Sumatran orangutans (Dierenfeld 1997; van Schaik *et al.* 2009b). This may allow Bornean orangutans to physiologically buffer against starvation (Knott 1998; Morrogh-Bernard *et al.* 2009; van Schaik *et al.* 2009b; Isler 2014). Following this, it was predicted that *P. p. morio* is most susceptible to obesity (van Schaik *et al.* 2009b).

Another striking example of the broad phenotypic variation in orangutans was that Bornean orangutans exhibit a faster-paced life history than those on Sumatra (van Schaik *et al.* 2009b). Along the west–east gradient, interbirth intervals (Wich *et al.* 2009a), time until association with the mother drops significantly, and age at first birth (van Noordwijk *et al.* 2009; C. P. van Schaik 2010, unpublished data) decrease. In addition, Sumatran orangutans are much more sociable and exhibit a higher social tolerance than Bornean orangutans (van Schaik 1999; van Schaik 2004; Knott *et al.* 2008; Mitra Setia *et al.* 2009; Weingrill *et al.* 2011). Probably linked to this and their larger brains, the cultural repertoire of Sumatran orangutans is larger and multiple complex innovations have been documented that are rare to absent within Bornean orangutans (van Schaik 2004; van Schaik *et al.* 2009a; Krützen *et al.* 2011). There is also a clear difference between the islands with regard to male sociosexual strategies and mating behavior (Utami *et al.* 2002; Utami-Atmoko *et al.* 2009; Dunkel *et al.* 2013).

Table 1. A selection of observed differences between orangutan taxa. Adapted from van Schaik *et al.* (2009b) and van Schaik (2013). For a complete overview see source studies. The species/subspecies are arranged from west to east in the same direction as the ecological gradient.

	Sumatra *P. abelii*	**Borneo** *P. p. wurmbii*	**Borneo** *P. p. morio*
Habitat			
Forest productivity	Higher	Lower	Lower
Impact of mast fruiting	Less	More	Most?
Tigers	Present	Absent	Absent
Morphology			
Average brain size (cc)	388	374	364
Mandibles	Gracile	Robust	Very robust
Tooth enamel	Thinner	Thicker	Thicker
Physiology			
Ketone bodies in urine	Very rare	Rare	?
Behavioral Ecology			
Variation in fruit intake	Low	Higher	Highest
Reliance on non-fruit fallbacks	Very rare	Common	Commonest
Mean insectivory (% feeding time)	*ca* 11%	*ca* 6%	*ca* 1.5%
Female daily travel distance (m)	*ca* 820	*ca* 760	*ca* 230
Number of day nests build/day	*ca* 0.8	*ca* 0.4	*ca* 0.05
Sensitivity to logging	High	Lower	Lowest
Population density	Higher	Usually Lower	Among lowest
Social organization			
Sociability	Highest	Lower	Lower
Susceptibility to social stress	Lower	Higher	Higher
Cultural repertoire	Large	Smaller	Small
Presence of complex innovations	Multiple	Rare	Absent
Earshot associations (fl. male-female)	Present	Absent	Absent
Male developmental arrest	Present	Weak	Absent?
Presence of forced matings	Rather low	High	High
Life history			
Interbirth intervals (mean, years)	8.75	7.70	6.10
Age at first birth	15–16	13–15	less than 13
Reduced association with mother	From *ca*. 10 years	From *ca*. 6 years	From *ca.* 6 years

1.3 The genomics revolution

Until recently, extensively examining the potential genetic basis of adaptations in natural populations has been beyond reach. Studying genetic targets of selection was methodologically limited to hypothesis-driven candidate gene approaches. In humans (and some other species) the examination of individual candidate genes with known phenotypic effect of the selected variant has yielded notable success for a few genes, including the lactose tolerance gene *LCT* (Bersaglieri *et al.* 2004), and genes that reduce malaria susceptibility such as *HBB* (Currat *et al.*, 2002; Ohashi *et al.,* 2004). However, such candidate gene approaches required existing knowledge about target genes, and were strongly restricted by the number of genes that could be sequenced.

The advent of high-throughput sequencing has transformed our ability to study the genetic basis of adaptive evolution (Storz 2005; Akey 2009) by enabling the generation of massive amounts of sequence data within a reasonable time and financial budget (Figure 3). This revolution in DNA sequencing technology provides an unprecedented opportunity to detect signatures of selection across the genome even without a prior knowledge of the associated phenotype, thus allows moving from hypothesis-driven to hypothesis-generating approaches (e.g. reviewed in Bank *et al.* 2014; Ellegren 2014; Pardo-Diaz *et al.* 2014). High-throughput sequencing has opened up new avenues in the study of evolutionary genetics and finally makes answers accessible to long-standing questions about genetic variation, adaptation and speciation (see above).

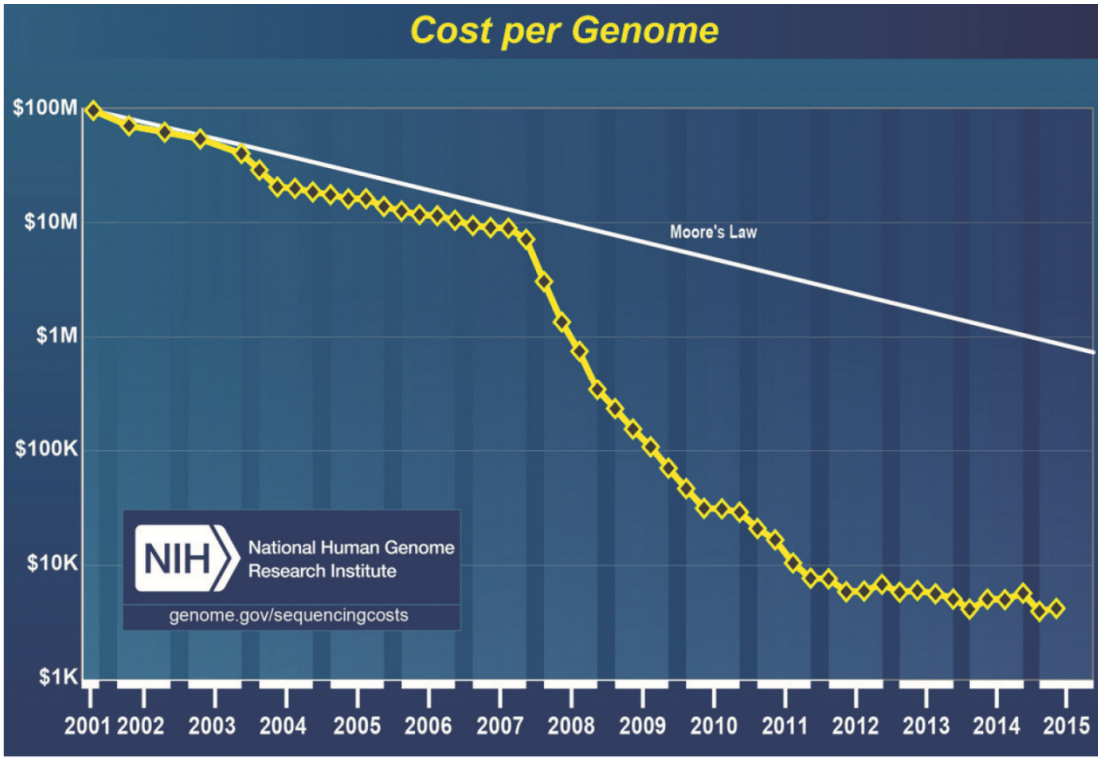


Figure 3. Overview of decrease of costs to sequence a human genome over the past decade. Taken from https://www.genome.gov/sequencingcosts/

1.4 Aims and outline of the dissertation

This dissertation seeks to contribute novel insights into the evolutionary history of the Asian great apes, by taking into account the environmental processes shaping it and the genomic basis underlying adaptations. I take advantage of the revolution in DNA sequencing technologies to tackle the evolutionary history of *Pongo* by applying a population genomics approach. A detailed understanding of how different evolutionary forces have impacted the patterns of genetic variation within and between orangutan species may ultimately provide insights into important processes during hominid evolution and will improve our understanding of evolution in general.

The role of selection and the genetic basis of local adaptations have been studied extensively in humans (e.g. Akey 2009; Pickrell *et al.* 2009; Pritchard *et al.* 2010; Hernandez *et al.* 2011; Grossman *et al.* 2013). However, similar attempts within genera or on the species level of non-human great apes are rare (but see McManus *et al.* 2015; Xue *et al.* 2015), and the required genomic sequence data are just beginning to emerge (Locke *et al.* 2011; Prado-Martinez *et al.* 2013; Scally *et al.* 2013; Xue *et al.* 2015). To date, orangutan adaptive evolution has only been investigated broadly on the level of the genus (e.g. Kosiol *et al.* 2008; Enard *et al.* 2010; Locke *et al.* 2011; Ma *et al.* 2013). However, our detailed knowledge about orangutan biology and its geographic variation offers a unique opportunity to study the genetic basis underlying adaptive phenotypic variation both within and between orangutan species.

Previous genetic work, in some of which I was actively involved during my Ph.D. candidacy (Arora *et al.* 2012; Nater *et al.* 2013; Nater *et al.* 2015), provided important insights into the complex population history and phylogeography of orangutans, but also yielded conflicting results. Central aspects of the evolutionary history of orangutans remain poorly understood such as the speciation process of Bornean and Sumatran orangutans. All previous studies of wild orangutan populations were based on a small number of classical genetic markers, such as microsatellites or short mtDNA sequences, which could be applied to non-invasively collected samples. However, to understand orangutan evolutionary history comprehensively, it is paramount to study their demographic history, population structure, and phylogeographic patters using genome-wide sequence data. In two recent sequencing efforts, Locke *et al.* (2011) and Prado-Martinez *et al.* (2013) sequenced whole genomes of ten orangutans each. These data represent highly valuable resources for orangutan genomics. However, inferences were hampered by the fact that all study individuals were captive zoo orangutans with unknown population provenance (though most individuals were wild-caught), which is limiting the perspective for population genomic analyses of wild populations.

This dissertation is aimed at (i) reconstructing the evolutionary history of the genus *Pongo* using genome-wide data from orangutans with known provenance across the genus' extant geographic range, and (ii) identifying potential genomic signatures of local adaptations. In order to achieve my goals related to orangutan biology, I focused in the first part of this

dissertation on methodological aspects, i.e. how to generate suitable population genomic data. This was because despite the recent advances in DNA sequencing technology, generating genomic sequence data from many individuals of a population still poses significant challenges in the laboratory and during bioinformatical analyses (Helyar *et al.* 2011; Nielsen *et al.* 2011; Steiner *et al.* 2013; Perry 2014), particularly for taxa with large genomes such as great apes. In the second part of this dissertation I studied the demographic history, population structure, phylogeographic patterns, and potential genetic local adaptations of orangutans, based on a unique dataset of autosomal and sex-specific genome data. In the following section, I provide a brief outline about the work performed within the framework of this dissertation.

Thesis outline

The quest for the Y

Genetic data from the male-specific region of the Y chromosome (MSY) represents an essential complement to maternally and biparentally inherited genetic markers, and is critical for studying male-specific evolutionary processes (Prugnolle & de Meeus 2002; Handley & Perrin 2007a). Because sex-biased dispersal may strongly impact the genetic makeup of natural populations, a comprehensive understanding of a species' evolutionary history necessitates the inclusion of sex-specific genetic markers, i.e. mitochondrial and Y-chromosomal loci in mammals. Mitochondrial markers are easily accessible and have been applied successfully in population genetics since decades. The development of useful MSY-specific single-copied markers, however, is technically challenging due to the highly complex architecture of the Y chromosome. Thus, male-specific genetic data have remained elusive for most mammalian species.

In **Chapter 2**, published as an invited technical review in Molecular Ecology Resources (Greminger *et al.* 2010), I present an overview of the current methodological strategies applied to developing MSY-specific genetic markers in non-model species and their practical feasibility and limitations. Furthermore, I describe strategies with future prospects with regard to the advent of high-throughput sequencing.

In **Chapter 5,** submitted to Systematic Biology (Greminger *et al.*, submitted), I present a novel bioinformatics strategy to extract MSY-specific single-copy sequences from whole-genome sequencing data. This approach allowed us for the first time to comparatively trace both the male- and female-specific evolutionary history on a genomic level in a non-human great ape (but see Xue *et al.* 2015). I also identified a large number of MSY-specific microsatellite markers and single-nucleotide polymorphisms (SNPs) which serve as a valuable resource for future studies of non-invasively sampled wild orangutans. To my knowledge, comparable Y chromosome sequencing in non-human mammals has only been achieved for horses (Wallner *et al.* 2013; Schubert *et al.* 2014), Mountain gorillas (Xue *et al.* 2015), as well as

polar and brown bears (Bidon *et al.* 2014). Our results of different evolutionary trajectories of males and females in orangutans demonstrate the great importance and power of genomic MSY-specific data for the comprehensive understanding of a species' evolutionary history. I expect that the principle of my bioinformatics strategy will be widely applicable to other mammalian species. In fact, a highly similar strategy has very recently been applied to Mountain gorillas (Xue *et al.* 2015).

Reduced genome complexity sequencing

Despite advances in DNA sequencing technology, (re-)sequencing whole genomes of many samples still constitutes substantial financial and computational effort, although it became more accessible at very recent times. Reduced genome complexity sequencing strategies (commonly known as RAD and RRL sequencing) offer great prospects for the generation of population genomic sequence data by allowing sampling only a fraction of the genome.

In **Chapter 3**, published in BMC Genomics (Greminger *et al.* 2014), I developed a novel protocol (named iRRL) for improved reduced genome complexity sequencing. Using this protocol, I generated iRRL data from the two populations at the extremes of the west–east gradient of variation of phenotypic traits in orangutans. The main strengths of my iRRL method are the very high genotyping-by-sequencing efficiency and reproducibility of genome complexity reduction among samples. My iRRL protocol is part of a growing suite of reduced complexity sequencing strategies that have transformed our ability to generate genomic data from natural populations.

Evaluation of SNP- and genotype calling

From a bioinformatical point of view, translating raw high-throughput sequencing data into high-quality SNP and genotype calls is challenging and requires many computational steps (Li *et al.* 2009; DePristo *et al.* 2011; Nielsen *et al.* 2011; Pabinger *et al.* 2013). Based on the iRRL data generated from two orangutan populations, I directly compared three commonly used SNP and genotype callers (**Chapter 3**) and obtained substantially different SNP datasets depending on the caller algorithm, sequencing depth and filtering criteria. These inconsistencies affected scans to detect selective sweeps (low overlap of identified putative sweeps) and will likely also exert undue influences on demographic inferences as implied by shifts in the allele-frequency spectra. Since the beginning of my Ph.D. candidacy, major advancements have been made in the development of sophisticated probabilistic algorithms for SNP and genotype calling (Van der Auwera *et al.* 2013; Li 2014). Nevertheless, accurate und unbiased SNP and genotype calling still remains a challenge, in particular for low or medium coverage reduced genome complexity sequencing data of non-model organisms. For this type of data, it is usually not yet possible to apply machine learning algorithms for variant quality score recalibrations (McKenna *et al.* 2010; DePristo *et al.* 2011) as commonly done for whole genome sequencing data.

Whole-genome sequencing

Despite the proven usefulness of reduced genome complexity sequencing (e.g. Hohenlohe *et al.* 2010; Stölting *et al.* 2013), data obtained in this manner face several limitations with respect to certain biological questions, which necessitate the use of whole-genome data. For instance, many modeling approaches to infer demographic history (e.g. Li & Durbin 2011; Harris & Nielsen 2013) require whole-genome data. Moreover, scans to detect signals of natural selection greatly profit from increased power, specificity, and resolution if based on whole-genome data. Only with complete genome information, we can make use of the full spectrum of statistical tests, as well as actually pinpoint the genes and functional SNPs involved in local adaptation. Thus, to pursue the main goals of this dissertation, we decided to put our emphasis on a large collaborative effort to sequence whole genomes of 17 wild-born orangutans with good population provenance to medium–high coverage (**Chapter 4**). Samples subjected to whole-genome sequencing were carefully selected in order to complement previous sequencing efforts (Locke *et al.* 2011; Prado-Martinez *et al.* 2013), thereby achieving a complete representation of the entire extant geographic range of the genus *Pongo*. The inclusion of the 20 previously sequenced individuals without reported provenance (Locke *et al.* 2011; Prado-Martinez *et al.* 2013) was made possible by our detailed knowledge of orangutan phylogeography and population structure based on classical genetic markers (Chapter 3; Arora *et al.* 2010; Nater *et al.* 2011; Nietlisbach *et al.* 2012; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015), providing a *hitherto* unprecedented opportunity to identify the natal population of individuals retrospectively.

This unique dataset of orangutan whole-genome sequencing data constituted the fundament of the analytical work carried out in the **Chapters 4–6**. **Chapters 4** and **6** will be published together with additional analyses, for which we have extended our collaborative network, in a main integrative paper (Greminger, Nater *et al.*, *in prep*). **Chapter 5** has been submitted to Systematic Biology (Greminger *et al.*, submitted).

Demographic history and population structure

In **Chapter 4**, I investigated the demographic history of the genus *Pongo* and the geographic structure of autosomal genetic diversity. I found that the speciation of Bornean and Sumatran orangutans has been a gradual process over several hundred thousand years, heavily influenced by recurrent climate changes in Sundaland. My findings also revealed that Bornean and Sumatran orangutans were affected differently by the Pleistocene climate oscillations. While climate changes had a major impact on the evolutionary history of Bornean orangutans, likely causing repeated bottlenecks and a long-term population decline, Sumatran orangutans were much less affected and experienced a remarkably stable population history and structure throughout the Pleistocene. Only recently, they also faced a drastic population decline, likely caused by the Toba supereruption ~73 ka and prehistoric hunting by early hunter-gatherers. The former adds to the highly controversial discussion about the consequences of the Toba supereruption by providing, to my knowledge, the first

direct evidence of a strong regional impact of the supereruption on a large mammal. The findings presented in this chapter also have important ramifications for orangutan conservation and taxonomy, in particular with respect to the Batang Toru population, the only extant Sumatran orangutans south of Lake Toba.

Sex-specific phylogeography

In **Chapter 5**, I focused on the sex-specific evolutionary histories of orangutans. Analyzing large-scale MSY sequence data (outlined above) and complete mitochondrial genomes, I found that orangutan evolutionary history is not only a tale of two islands, but also one of two sexes. Males and females exhibited strikingly distinct population histories and phylogeographic patterns, owing to high levels of male-biased dispersal and strict female philopatry in orangutans. The results from the mitochondrial genomes further confirmed previous findings of a common late Pleistocene rainforest refugium of Bornean orangutans (Arora *et al.* 2010; Nater *et al.* 2011) as well as an extremely deep split of Sumatran orangutans to the north and to the south of Lake Toba (Arora *et al.* 2010; Nater *et al.* 2011). The genomic MSY data also shed light into the long-lasting debate when male-mediated gene flow ceased between Borneo and Sumatra (Harrison *et al.* 2006; Kanthaswamy *et al.* 2006; Steiper 2006; Locke *et al.* 2011; Nater *et al.* 2011; Nater *et al.* 2015), by revealing that the two species likely have been reproductively isolated for considerably longer time than proposed previously. The results presented in this chapter further suggest that different evolutionary forces might act on the MSY in the two orangutan species, probably linked to extensive reproductive skew among Sumatran males.

Genomic signatures of local adaptation

In **Chapter 6**, I present the first whole-genome scans for positive selection within the genus *Pongo* to study the genetic basis of local adaptations. Using a combination of approaches to detect signatures of positive selection, including window-based genome scans to identify putative hard sweeps, I identified strong candidate genes and functional SNPs potentially associated with the observed variation in phenotypic traits in orangutans (van Schaik *et al.* 2009b). In Bornean orangutans, I found for instance signals of potential adaptation pertaining to energy storage (i.e. adipose tissue) metabolism, in congruence with their greater ability to deposit large fat storages. I also identified several candidate genes and biological processes related to neurogenesis, which is in line with the smaller brain size of Bornean orangutans. In contrast, in Sumatran orangutans, I found for example signatures of potential adaptive evolution of genes related to learning, adult brain plasticity, and the oxytocin pathway. I hypothesize that selective changes in these genes may provide Sumatran orangutans with a framework allowing for extended behavioral plasticity, as mirrored in their larger and more complex cultural repertoire and their higher sociability. Overall, the results of this chapter suggest that both orangutan species experienced very different adaptive evolutionary histories and that at least some of the striking geographic variation in orangutan phenotypic

traits (van Schaik *et al.* 2009b; Wich *et al.* 2009b) may indeed represent genetic local adaptations.

Chapter 2

The quest for Y-chromosomal markers – methodological strategies for mammalian non-model organisms

Maja P. Greminger[1], Michael Krützen[1], Claude Schelling[2], Aldona Pienkowska–Schelling[3] and Peter Wandeler[4]

[1]Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

[2]Animal Genetics Group, Vetsuisse-Faculty Zurich, University of Zurich, Zurich, Switzerland

[3]Department of Animal Sciences, Federal Institute of Technology Zurich, Switzerland

[4]Zoological Museum, University of Zurich, Zurich, Switzerland

2.1 Abstract

Tracing maternal and paternal lineages independently to explore breeding systems and dispersal strategies in natural populations has been high on the wish list of evolutionary biologists. Because males are the heterogametic sex in mammals, such sex-specific patterns can be indirectly observed when Y chromosome polymorphism is combined with mitochondrial sequence information. Over the past decade, Y-chromosomal markers applied to human populations have revealed remarkable differences in the demographic history and behaviour between the sexes. However, with a few exceptions, genetic data tracing the paternal line are lacking in most other mammalian species. This deficit can be attributed to the difficulty of developing Y-specific genetic markers in non-model organisms and the general low levels of polymorphisms observed on the Y chromosome. Here, we present an overview of the currently employed strategies for developing paternal markers in mammals. Moreover, we review the practical feasibility and requirements of various methodological strategies and highlight their future prospects when combined with new molecular techniques such as next generation sequencing.

2.2 Introduction

Why Y?

Females and males do not pass on their genes equally to the next generation in space and time. In birds for example, dispersal is mainly female-biased and males are philopatric, resulting in a higher number of female genes being exchanged between populations. Such sex-biased dispersal has profound consequences for the genetic diversity and genetic makeup of natural populations (Handley & Perrin 2007b). Information on sex specific differences in dispersal patterns gathered using autosomal genetic markers is limited to one generation, given that recombination will obscure independent maternal and paternal lineages in the next generation (reviewed in Goudet *et al.* 2002; Prugnolle & de Meeus 2002). Moreover, sex-biased dispersal is currently still assessed qualitatively rather than quantitatively (Petit *et al.* 2002). As a consequence, there remains a considerable gap between empirical data and our theoretical understanding of dispersal strategies in relation to different mating systems and ecological constraints (Handley & Perrin 2007b).

The independent demographic population history either sex can be investigated by contrasting chromosome polymorphisms of the heterogametic sex with mitochondrial DNA (mtDNA) sequence variation, provided that maternally inherited mtDNA is passed on in the homogametic sex. In mammals, for instance, males are the heterogametic sex and thus inherit large portions of the Y-chromosome without recombination from their father. Y-specific haplotype data can therefore be applied as paternal analogues to mtDNA. Yet with the exception of primates (Erler *et al.* 2004; Eriksson *et al.* 2006; Douadi *et al.* 2007a), only a few studies report Y chromosome variation within and between natural populations (e.g. Sundqvist *et al.* 2001; MacDonald *et al.* 2006; Hailer & Leonard 2008). In humans, numerous Y-chromosomal markers have been discovered over the last decade (Kayser *et al.* 2004), and as such enabled the collection of comprehensive sets of paternal data. For example, comparisons between Y-variation with mitochondrial sequence information alongside autosomal data provide independent evidence for a single African origin of modern humans (Ke *et al.* 2001). Evidence for higher female migration rates among humans (Seielstad *et al.* 1998) and significant sex-specific differences of movements between social ranks by members of different Hindu castes for marital purposes (Bamshad *et al.* 1998) have been reported.

Despite these advances, the question of why there are so few studies contrasting maternal and paternal lineages remains. This is remarkable because of the apparent potential to explore breeding systems and dispersal strategies in animal populations. Most likely, the answer lies in the difficulty of discovering polymorphic Y chromosome-specific markers in natural populations. In order to be useful for most biological applications, these markers need to fulfil three criteria: specificity to the heterogametic sex, single locus amplification, and sufficient levels of polymorphism.

The combination of the distinctive architecture of the Y chromosome and low levels of genetic variation require a substantial methodological effort to develop markers meeting all three criteria. We will provide a brief overview of the Y chromosome architecture and potential reasons for its low genetic variation. This is followed by a review of the current methodological strategies applied to discovering Y-polymorphisms in non-model organisms, *i.e.* in species with little or no genomic information. Additionally, we highlight strategies with future prospects when combined with new molecular techniques.

Throughout this review, we will focus on studies concerned with developing Y-specific markers and on those inferring Y chromosome variation between and within populations of mammals. Nonetheless, the described methodological strategies can be analogically applied to discover other sex-chromosome-specific markers in any heterogametic species. Obviously, however, in female heterogametic (ZW) systems (*e.g.* birds, Lepidoptera), it will not be possible to independently contrast maternal and paternal lineages since both the W chromosome and mtDNA are maternally inherited.

The different architecture of the Y chromosome

The principle of chromosomal sex determination has evolved independently among taxa (Bull, 1984). The mammalian Y chromosome possesses only a limited number of active genes primarily responsible for spermatogenesis and male determination (Lahn & Page 1999). Because the X and Y mammalian sex chromosomes share a common ancestor, widespread sequence homologies can be found. In addition, the Y chromosome has acquired autosomal sequences by retrotransposition or gene conversion from other parts of the genome and vice versa (Steinemann & Steinemann 1992; Skaletsky *et al.* 2003b; Handley & Perrin 2006). However, our current knowledge of detailed Y chromosome architecture in mammals is restricted to humans.

In humans, only the pseudoautosomal region, comprising a small portion at either end of the Y chromosome, recombines with the X chromosome during meiosis (Graves *et al.* 1998). The remainder of the chromosome, called the male-specific region (MSY), consists of several blocks of highly palindromic heterochromatic sequences and a smaller euchromatic part (Skaletsky *et al.* 2003b). In contrast to dense heterochromatin, which mainly consists of non-transcribed repetitive satellite sequences, euchromatin decondenses during interphase and contains, among others, transcribed genes. Euchromatic DNA of the Y chromosome can be further subdivided into three different sequence classes: ampliconic segments, X-transposed, and X-degenerate sequences (Skaletsky *et al.* 2003b). Ampliconic segments are large tandemly repeated palindromic units showing high levels of intrachromosomal sequence identity. This sequence class may contain several copies of a single gene. The X-transposed sequences are a result of a large X- to Y- transposition that occurred after the divergence of the human and the chimpanzee lineage. Therefore, these are unique to the human Y chromosome. Finally, the X-degenerate sequence class represents relics of the ancient autosomes from which the X and Y chromosome are thought to have evolved, and comprises

single-copy genes homologues of the X-linked genes. Only paternal markers derived from loci within these X-degenerated sequences should instantly fulfil the criteria of male-specificity and single-locus amplification.

Undoubtedly, the development of Y-specific markers is compromised by the distinct architecture of the Y chromosome. Incidentally, this limitation is also reflected by the fact that the Y chromosome is neglected in most mammalian genome sequencing projects due to difficulties in generating sequence data and aligning contigs (Murphy *et al.* 2006). As such, females are used in most current genome projects. Inferences on size and structure of Y chromosomes are difficult as the Y-chromosomal architecture may differ significantly among and even within species (Waters *et al.* 2007): chromosome size, genetic structure including characteristics of X-degenerated sequences present on the Y chromosome, as well as content and relative location of the expressed genes may vary considerably (Figure 1, Kirsch *et al.* 2008).

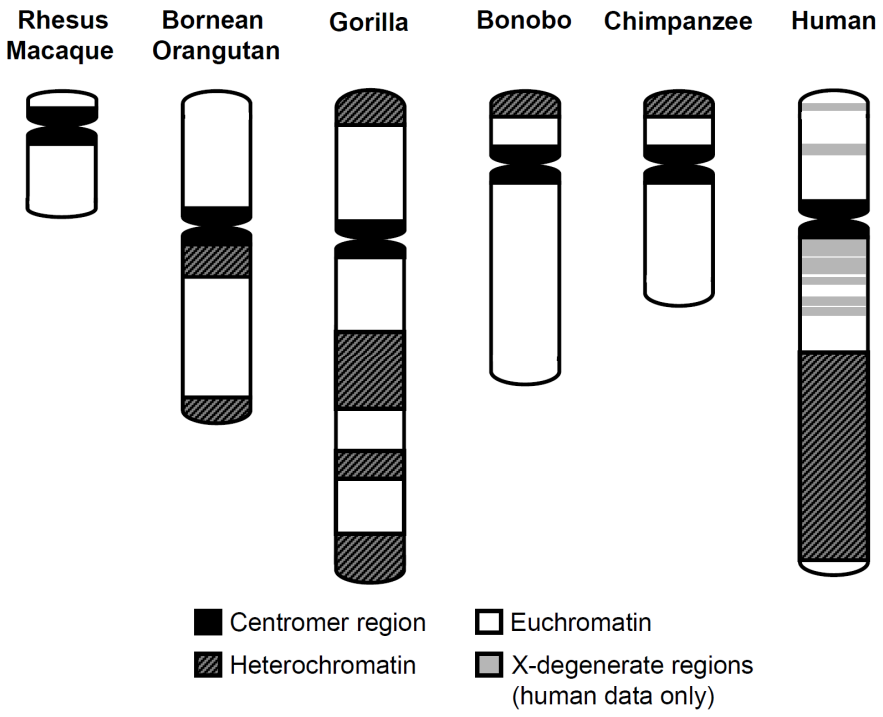


Figure 1. Dynamic evolution of the Y chromosome. Illustrated by a schematic comparison of great ape and macaque Y chromosomes (modified from Kirsch *et al.*, 2008) describing an evolutionary time span of ~23 million years. Y chromosomes differ substantially in size and structure, i.e. proportions and locations of sequence classes, among species. Only within a small fraction of the Y, the X-degenerate regions, the development of single-locus and male-specific genetic makers is likely to be successful. Note that detailed structural data are only available for the human Y chromosome (Skaletsky *et al.*, 2003).

Expected low levels of genetic variation

The haploid nature of the Y chromosome has important ramifications for the level of genetic diversity of Y sequences. Due to the general lack of recombination, new mutations are the only evolutionary force increasing DNA variation in the MSY. Mutation rates in the MSY are higher than in the remainder of the genome due to DNA replication errors during gametogenesis, in particular in organisms with longer generation times (reviewed in Makova & Li 2002; Goetting-Minesky & Makova 2006). However, empirical evidence for Y-linked and autosomal microsatellites shows no significant difference between their average mutation rates in humans (Heyer *et al.* 1997; Gusmao *et al.* 2005). Hence, additional data from other taxa are needed in order to address this question.

Despite the higher mutation rate, lower levels of nucleotide diversity on the Y chromosome relative to the rest of the nuclear genome can be expected with several major mechanisms being responsible. Under the assumption of a balanced sex ratio and equal variance in reproductive success among males and females, the effective population size (*Ne*) of the Y chromosome and the mitochondrion is identical, but equivalent to only one-quarter of that of the autosomes. Therefore, the genetic diversity of Y chromosome and mtDNA are particularly sensitive to demographic events such as population bottlenecks. In addition, the typical mammalian mating system is polygynous, where a few males father a disproportional fraction of the offspring (Greenwood 1980). This imposes a higher variance in reproductive success in males and consequently leads to a reduction in N_e of the paternal lineage (Caballero 1995). Finally, the MSY behaves like a single locus and as such is thought to be susceptible to the influence of selective forces acting on the Y chromosome reducing its genetic variation (Charlesworth & Charlesworth 2000).

Low levels of Y-polymorphism have been recorded in species of different mammalian orders (e.g. Hellborg & Ellegren 2004; but see Andres *et al.* 2008). Therefore, a substantial sequencing effort is often needed to detect informative SNPs. As a result, researchers have resorted to identifying genetic markers on the Y chromosome with higher mutation rates, *e.g.* microsatellites (Wallner *et al.* 2004; Handley & Perrin 2006; Luo *et al.* 2007). While the identification of polymorphic markers located on the MSY can be difficult itself, there is an advantage in that only a few polymorphic markers combined to haplotypes are sufficient to describe paternal genetic diversity. For example, in humans only seven polymorphic microsatellites are sufficient to depict the vast majority of the global haplotype diversity (Kayser & Sajantila 2001).

2.3 Current methodological strategies

Although the number of Y-specific and polymorphic genetic markers across the mammalian order published to date has been rather small (Table 1), different methodological strategies have been applied. Current methods range from the simple cross-amplification using primers

designed in conserved MSY regions and exploring of intronic sequence variation to complex multileveled development strategies and Y-specific enriched microsatellite libraries. All employed strategies can be divided into two groups: strategies suitable for screening a pool of several individuals for sequence polymorphisms (SNPs, indels and microsatellites) and strategies exploring Y-specific material from one individual for microsatellite repeat motifs (Figure 2).

Conserved exonic Y-sequences

Although comprehensive Y-sequence data are lacking in non-model organisms, exonic sequence data of MSY genes of an increasing number of species can be found in public databases. This information can be used to design exonic PCR primers flanking MSY gene introns, referred to as "Y chromosome conserved anchored tagged sequences" (YCATS, Figure 2). Hellborg and Ellegren (2003) established this approach for the Y chromosome based on the CATS concept (Lyons *et al.* 1997). YCATS primers are assumed to provide access to intronic noncoding DNA, which can be screened for genetic variability. The conserved nature of the exonic primers facilitates the amplification across a range of related species. Based on human-mouse Y sequence alignments, Hellborg and Ellegren (2003) developed a YCATS marker system which has been frequently applied in studies investigating Y chromosome polymorphism in a wide range of mammalian species (Table 1). A similar approach has been applied by Andrés *et al.* (2008) to develop a microarray system for Y-chromosomal SNPs in the chimpanzee (*Pan troglodytes*). In this particular study, PCR primers based on human and chimpanzee sequence alignments were applied to sequence 9.1 kb of different MSY gene introns.

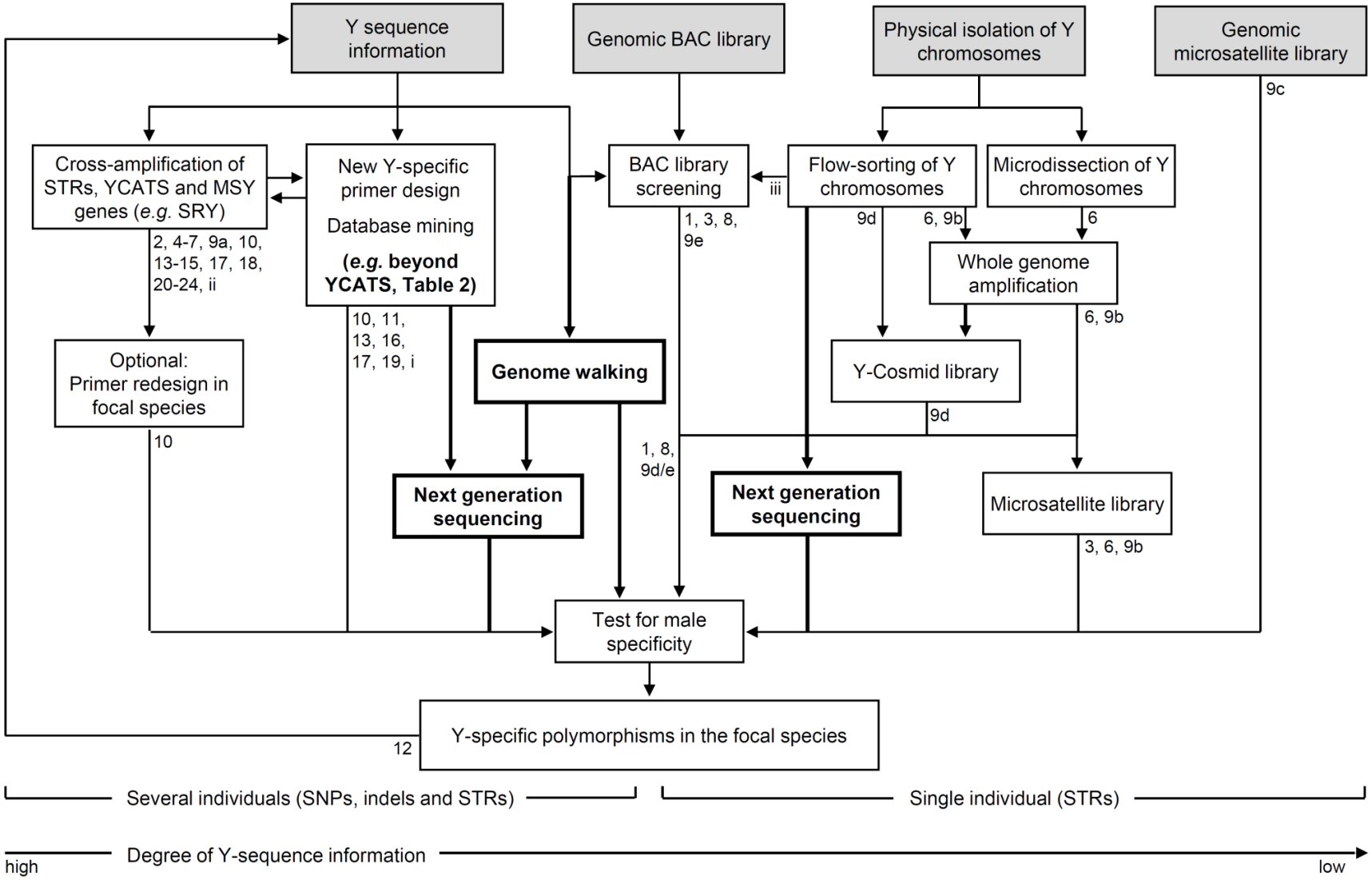

Y sequence information
Genomic BAC library
Physical isolation of Y chromosomes
Genomic microsatellite library
9c
Cross-amplification of STRs, YCATS and MSY genes (e.g. SRY)
New Y-specific primer design
Database mining
(e.g. beyond YCATS, Table 2)
BAC library screening
iii
Flow-sorting of Y chromosomes
Microdissection of Y chromosomes
2, 4-7, 9a, 10, 13-15, 17, 18, 20-24, ii
10, 11, 13, 16, 17, 19, i
1, 3, 8, 9e
9d
6, 9b
6
Whole genome amplification
6, 9b
Optional: Primer redesign in focal species
Genome walking
Y-Cosmid library
9d
10
1, 8, 9d/e
Next generation sequencing
Next generation sequencing
Microsatellite library
3, 6, 9b
Test for male specificity
Y-specific polymorphisms in the focal species
12
Several individuals (SNPs, indels and STRs)
Single individual (STRs)
high
Degree of Y-sequence information
low

Figure 2. Overview of current and emerging strategies for obtaining Y-specific and polymorphic markers. Based on their properties, strategies can be divided into two categories: a) strategies screening a pool of several individuals for sequence polymorphisms (SNPs, indels and STRs); and strategies exploring Y-specific material from a single individual for microsatellite repeat motifs. Different source materials are presented in grey boxes. Bold pathways indicate emerging strategies. The numbers below the boxes refer to studies listed in Table 1 which have applied the particular strategy. Studies not listed in Table 1: i, Hellborg & Ellegren, 2003; ii, Erler *et al.*, 2004; iii, Sankovic *et al.*, 2006. Abbreviations: YCATS, Y chromosome conserved anchored tagged sequence; MSY, male-specific region of the Y; STRs, microsatellite markers; SNPs, single nucleotide polymorphisms; BAC, bacterial artificial chromosome. Note that the required amount of initial Y sequence information is decreasing gradually from the left to the right.

Table 1 Selected studies with the aim to develop or apply Y-specific genetic markers in non-human mammals. An overview of the employed methodological strategies is given in Figure 2. Abbreviations: Dev, development of novel genetic markers; STRs, microsatellite loci; SNPs, single nucleotide polymorphisms; NSTRs, number (no.) of microsatellite loci x/y/z (x= no. of obtained STRs, y= no. of male-specific and single-copied amplified STRs, z= no. of polymorphic STRs); NA, total number of alleles across all loci; TSL, total sequence length [kb]; NPLS, total number of polymorphic sites; NIND: total number of individuals; NHAP, total number of haplotypes; ISV, intra-species variation; IPV, intra-population variation; RF, reference number in Figure 2; SRY, sex determining region Y; N/A or (?), not available information; dash (-), has not been investigated. Note that phylogenetic studies were not considered.

			STRs		SNPs		Y diversity					
Species	**Methodological strategies (Figure 2)**	**Dev**	N_{STRs}	N_A	TSL	N_{PLS}	N_{IND}	N_{HAP}	ISV	IPV	**RF**	**Reference**
Marsupials												
Tammar wallaby (*Macropus eugenii*)	Genomic BAC library	Yes	20/10/4	19	-	-	22	11	Yes	N/A	1	MacDonald *et al.* 2006
Insectivores												
Common shrew (*Sorex araneus*)	Cross-amplification of STRs	No	1/1/1	14	-	-	70	N/A	Yes	Yes	2	Lugon-Moulin & Hausser 2002
Greater white-toothed shrew (*Crocidura russula*)	Genomic BAC library, STR library	Yes	8/1/1	8	-	-	81	N/A	Yes	N/A	3	Handley & Perrin 2006
	Cross-amplification of YCATS & SRY	No	-	-	1.67	10	49	N/A	Yes	N/A	4	Handley *et al.* 2006a
Valais shrew (*Sorex antinorii*)	Cross-amplification of STRs	No	1/1/1	19	-	-	113	N/A	Yes	Yes	5	Yannic *et al.* 2008
Ungulates												
Bottlenose dolphin (*Tursiops aduncus*)	Physical isolation of Y chromosomes, STR library, cross-amplification of YCATS	No	37/11/0	-	1.21	4	28	4	Yes	Yes	6	Greminger 2007
Cattle (*Bos taurus*)	Cross-amplification of YCATS & SRY	No	-	-	3.50	0	10	N/A	No	No	7	Hellborg & Ellegren 2004
Domestic horse (*Equus caballus*)	Genomic BAC library	Yes	11/5/0	1	-	-	49	1	No	No	8	Wallner *et al.* 2004
Reindeer (*Rangifer tarandus*)	Cross-amplification of YCATS & SRY	No	-	-	0.70	0	20	N/A	No	No	7	Hellborg & Ellegren 2004
Carnivores												
Asian gold cat (*Pardofelis temmincki*)	Cross-amplification of YCATS & STRs	No	1/1/0	1	2.17	1	29	2	Yes	N/A	9a	Luo *et al.* 2007

Table 1 (Continued)

Species	Methodological strategies (Figure 2)	Dev	STRs		SNPs		Y diversity				RF	Reference
			N_{STRs}	N_A	TSL	N_{PLS}	N_{IND}	N_{HAP}	ISV	IPV		
Asian leopard cat (*Prinailurus bengalensis*)	Cross-amplification of YCATS & STRs	No	41/4/2	13	2.17	7	10/83	6**	Yes	N/A	9a	Luo *et al.* 2007
	Physical isolation of Y chromosomes, STR library	Yes	29/0/-	-	-	-	-	-	-	-	9b	Luo *et al.* 2007
Coyote (*Canis latrans*)	Cross-amplification of STRs, new Y-specific primer design	Yes	4/4/4	22	-	-	70	26	Yes	Yes	10	Hailer & Leonard 2008
Domestic cat (*Felis catus*)	Genomic STR library	Yes	380/0/-	-	-	-	-	-	-	-	9c	Luo *et al.* 2007
	Physical isolation of Y chromosomes, cosmid library	Yes	34/2/0	1	-	-	10	1	No	No	9d	Luo *et al.* 2007
	Genomic BAC library	Yes	18/3/0	1	-	-	10	1	No	No	9e	Luo *et al.* 2007
	Cross-amplification of YCATS & STRs	No	1/1/0	1	2.17	1	10	1	No	No	9a	Luo *et al.* 2007
Domestic dog *Canis familiaris*)	Data base mining, new Y-specific primer design	Yes	-	-	14.4	14	10	9	Yes	Yes	11	Natanaelsson *et al.* 2006
	Y-specific STRs in focal species	No	1/1/1	8	-	-	38	N/A	Yes	N/A	12	Vila *et al.*
Fishing cat (*Prionailurus viverrinus*)	Cross-amplification of YCATS & STR	No	1/1/1	3	2.17	4	24	2**	Yes	N/A	9a	Luo *et al.* 2007
Grey wolf (*Canis lupus*)	Cross-amplification of STRs, new Y-specific primer design	Yes	4/4/4	20	-	-	100	17	Yes	N/A	13	Sundqvist *et al.* 2001
	Cross-amplification of STRs	No	4/4/4	N/A	-	-	112	4	Yes	No	14	Sundqvist *et al.* 2006
	Cross-amplification STRs	No	4/4/4	N/A	-	-	202	19	Yes	Yes	15	Musiani *et al.* 2007
	Cross-amplification of YCATS & SRY	No	-	-	1.60	2	36	N/A	Yes	N/A	7	Hellborg & Ellegren 2004
Leopard (*Panthera pardus*)	Cross-amplification of YCATS & STRs	No	1/1/0	0	2.17	3	75	3	Yes	N/A	9a	Luo *et al.* 2007
Lynx (*Lynx lynx*)	Cross-amplification of YCATS & SRY	No	-	-	2.00	0	40	N/A	No	No	7	Hellborg & Ellegren 2004
Marbled cat (*Pardofelis marmorata*)	Cross-amplification of YCATS & STRs	No	1/1/1	2	2.17	4	8	2**	Yes	N/A	9a	Luo *et al.* 2007
Red Wolf (*Canis rufus*)	Cross-amplification of STRs, new Y-specific primer design	Yes	4/4/3	7	-	-	5	2	Yes	N/A	10	Hailer & Leonard 2008
Tiger (*Panthera tigris*)	Cross-amplification of YCATS & STRs	No	41/4/1	2	2.17	0	10/55	2	Yes	N/A	9a	Luo *et al.* 2007
	Physical isolation of Y chromosomes , STR library	Yes	14/0/-	-	-	-	-	-	-	-	9b	Luo *et al.* 2007

Table 1 (Continued)

Species	Methodological strategies (Figure 2)	Dev	STRs		SNPs		Y diversity				RF	Reference
			N_{STRs}	N_A	TSL	N_{PLS}	N_{IND}	N_{HAP}	ISV	IPV		
Rodents												
Field vole (*Microtus agrestis*)	Cross-amplification of YCATS & SRY	No	-	-	3.20	4	18	N/A	Yes	N/A	7	Hellborg & Ellegren 2004
Snow vole (*Chionomys nivalis*)	New Y-specific primer design	Yes	1/1/1	5	12.4	34	8	9	Yes	Yes	16	Wandeler & Camenisch in prep.
Primates												
Bonobo (*Pan paniscus*)	Cross-amplification of STS, new Y-specific primer design	No	-	-	2.78	4	7	3	Yes	N/A	17	Stone *et al.* 2002
	Cross-amplification of STRs	No	31/?/10	36	-	-	34	13	Yes	Yes	18	Eriksson *et al.* 2006
Chimpanzee (*Pan troglodytes*)	Data base mining, new Y-specific primer design, Y-intron sequencing	Yes	-	-	7.95	23	61	21	Yes	N/A	19	Andres *et al.* 2008
	Cross-amplification of STSs, new Y-specific primer design	No	-	-	2.78	19	101	10	Yes	N/A	17	Stone *et al.* 2002
Gorilla (*Gorilla gorilla gorilla*)	Cross-amplification of STRs	No	16/?/6	31	-	-	57	34	Yes	Yes	20	Douadi *et al.* 2007a
Hamadryas baboon (*Papio hamadryas*)	Cross-amplification of YCATS & STRs	No	7/4/2	4	3.60	0	97	1	Yes	Yes	21	Handley *et al.* 2006b
Japanese macaque (*Macaca fuscata*)	Cross-amplification of STRs	No	3/3/3	N/A	-	-	42	13	Yes	Yes	22	Kawamoto *et al.* 2008b
Long-tailed Macaque (*Macaca fascicularis*)	Cross-amplification of STRs	No	3/3/0	1	-	-	38	1	No	No	23	Kawamoto *et al.* 2008a
Moor macaque *(Macaca maura)*	Cross-amplification of STRs	No	1/1/1	7	-	-	14	7	Yes	Yes	24	Evans *et al.* 2001
Tonkean Macaque *(Macaca tonkeana)*	Cross-amplification of STRs	No	1/1/1	9	-	-	20	9	Yes	Yes	24	Evans *et al.* 2001

** only SNP data haplotypes

Cross-species amplification

Currently, the most straightforward strategy for obtaining Y-chromosomal data is to take advantage of Y-sequence information available from a closely related species. In this approach, PCR primers developed for one species will be used in an species of interest (Figure 2). To date, this approach has been used in the majority of studies investigating Y-specific polymorphisms (Table 1). Researchers working on primates and especially great apes have mostly benefited from this strategy, as the human genome project has facilitated the identification of a large number of Y-chromosomal microsatellite markers (Kayser *et al.* 2004). These markers have been tested for cross-species amplification in a broad range of primates (Erler *et al.* 2004) and applied in studies of bonobos (Pan paniscus, Stone *et al.* 2002; Eriksson *et al.* 2006), gorillas (Gorilla gorilla, Douadi *et al.* 2007a), hamadryas baboons (Papio hamadryas, Hammond *et al.* 2006) and macaques (Macaca fascicularis, Kawamoto *et al.* 2008a) among others. Similarly, polymorphic microsatellites described in the domestic dog (*Canis familiaris*) have been successfully amplified in wolves (Canis lupus, Sundqvist *et al.* 2001; Sundqvist *et al.* 2006; Musiani *et al.* 2007).

Although cross-species amplification has been the most popular approach so far, its success has been limited (Table 1). The major drawback of cross-species amplification is that the PCR amplification success is inversely related to the evolutionary distance between the species from which the loci have been isolated and the species to which the loci are being applied, due to the accumulation of mutations in the primer-binding sites (Primmer *et al.* 1996). These mismatches in the primer binding sites decrease the overall PCR efficiency and as such cause the frequently observed problem of unspecific binding. This problem is further exacerbated by the highly repetitive structure of the Y chromosome and its sequence homologies to the X (Figure 1). Obviously, this observation holds to a lesser extent for the amplification of the more conserved exonic YCATS primers in comparison to cross-amplification of microsatellites (Hellborg & Ellegren 2003). Importantly, these effects are enhanced using Y-chromosomal markers compared to autosomal loci due to the faster evolution of the Y chromosome (Erler *et al.* 2004). In addition, overall microsatellite polymorphism decreases with the phylogenetic distance to the source species (Primmer *et al.* 1996). Nonetheless, efficient Y-haplotyping by PCR of cross-species identified microsatellites can be achieved by re-designing species specific primers within the flanking region of the microsatellite.

Physical isolation of Y chromosomes

Most researchers working on non-model species face the problem of no or very limited Y-sequence information. The difficulty therefore is how to obtain and identify Y-chromosomal material for subsequent genetic marker development. One known strategy is to physically isolate the entire Y chromosomes or parts thereof from the focal species, using cytogenetic methods such as fluorescence-activated cell-sorting (FACS, Ferguson-Smith, 1995) or microdissection (Figure 2, Pienkowska-Schelling *et al.* 2005). Fluorescence-activated cell-

sorting relies on the discrete staining of metaphase chromosomes whereby the chromosome of interest is sorted according to its characteristic dye content using a flow cytometer (Bergstrom *et al.* 1998). Alternatively to FACS, Y chromosomes can be physically isolated by microdissection, *i.e.* by scraping stained chromosomes from metaphase spreads under an inverted microscope (Pienkowska-Schelling *et al.* 2005). While successful cell sorting can provide several hundred copies of the chromosome, the number of chromosomes retrieved by microdissection is limited. As a consequence, microdissection and any molecular work carried out thereafter require laboratory standards similar to working with highly diluted and degraded DNA extracted from museum samples Wandeler *et al.*, 2007). A prerequisite for both methods is the recognition of the Y chromosome, and as such good knowledge of the karyotype of the species of interest. Unfortunately, this information is frequently insufficient. Even in cases where karyotypes from a related taxon are available, the distinct identification of the Y chromosome is often difficult due to its dynamic evolution (Figure 1, Kirsch *et al.* 2008).

To our knowledge, only one study so far has reported a cytogenetic strategy for the development of Y-specific markers (Luo *et al.* 2007). Here, fluorescence-activated Y chromosomes from fibroblast cell cultures of the domestic cat (*Felis catus*) were isolated. Two different approaches were then applied to clone microsatellite markers. First, microsatellites were discovered through partial sequencing of a cosmid library established from the flow-sorted chromosomes. Second, the flow-sorted DNA was amplified through whole genome amplification (WGA) to obtain sufficient material for the establishment of a microsatellite-enriched library. Despite this substantial methodological effort, only 24 of 77 isolated microsatellites were male-specific and no marker was found to be polymorphic (Luo *et al.* 2007).

We deem the moderate success of Luo's study as typical. We know of several research groups, including ourselves, who have tried to develop Y-specific microsatellites by employing cytogenetic methods in a diverse range of mammalian species. To our knowledge, however, only a very small number of the developed markers fulfil the three criteria (*i.e*. specificity to the heterogametic sex, single-locus amplification and sufficient levels of polymorphism). The reasons for this sobering result could be multifaceted, but most likely lie in the unique architecture of the Y chromosome itself and the technical complexity of the isolation procedures. The palindromic structure plus the presence of autosomal sequence transpositions, as well as a relatively small proportion of the chromosome consisting of X-degenerated regions will lead to two major problems. Firstly, most discovered microsatellites will be present in multiple copies, leading to problems in designing locus-specific primer pairs. Secondly, most loci will not amplify male-specific. In both the microdissection and FACS approaches, Y chromosomes are often difficult to distinguish from X chromosomes and similarly sized autosomes. As a consequence, autosomal contamination of the physically isolated material can be expected and as such will decrease the likelihood of discovering male specific markers substantially. Moreover, the often required initial amplification of the

isolated and likely condensed DNA by WGA techniques can suffer from biased or non-specific amplification (Jiang *et al.* 2005).

Genomic libraries

An alternative approach to the physical isolation of Y chromosomes is to develop male-specific markers from total genomic DNA (Figure 2). An increasing number of genomic bacterial artificial chromosomes (BAC) libraries for different mammalian species are publicly available. They may serve as a source to discover Y-linked microsatellites or to obtain Y-specific sequence information. Clones containing Y-chromosomal sequences can be identified through screening of the BAC library with Y-specific probes such as SRY gene sequences (Luo *et al.* 2007), YCATS amplicons (Handley & Perrin 2006) or degenerate oligonucleotide primed PCR (DOP-PCR) products from microdissected or flow sorted Y chromosomes (Sankovic *et al.* 2006). Male-specific markers can subsequently be discovered by subcloning or shotgun sequencing of the selected clones (Wallner *et al.* 2004; MacDonald *et al.* 2006; Luo *et al.* 2007) or by constructing a Y-specific microsatellite-enriched library (Handley & Perrin 2006; Luo *et al.* 2007).

Nonetheless, this strategy proved to be only moderately successful in discovering male-specific microsatellites in different mammalian species. Although a number of microsatellite loci could be identified, only few appeared to be male-specific and even fewer were also polymorphic (Table 1). A potential reason for this might be that the BAC clones selected for the discovery of microsatellite motifs contained other sequences than from the X-degenerated regions. In addition, the length of the Y-specific BAC insert used for the construction of a microsatellite-enriched library might be a limiting factor. In order to maximize the amount of Y-specific template sequences, it is recommended to screen the genomic BAC library with amplicons of several single-copied MSY-specific genes located in the X-degenerated regions (Handley & Perrin 2006, Table 2). The use of DOP-PCR products as probes should be avoided as these hybridize randomly to Y chromosome sequences.

Finally an alternative strategy for the identification of Y-linked markers from total genomic DNA could be the screening of genomic microsatellite-enriched libraries. However, the success of such an approach is likely to be limited given the expected small proportion of enriched MSY-specific fragments relative to genomic fragments (Luo *et al.* 2007).

Table 2. Single-copied MSY-linked genes as targets for intron sequencing. Sequence data were taken from the NCBI database (http://www.ncbi.nlm.nih.gov). Gene abbreviations: SMCY, SMC (mouse) homologue Y (aliases Jarid1d, Kdm5d); DBY, dead box Y (aliases DDX3Y); UTY, ubiquitous TPR motif Y; EIF1AY, eukaryotic translation initiation factor 1A Y; UBE1Y, ubiquitin-activating enzyme E1 Y.

	Human (*Homo sapiens*)				**Chimpanzee (*Pan tryglodytes*)**				**Mouse (*Mus musculus*)**				
	SMCY	**DBY**	**UTY**	**EIF1AY**	**SMCY**	**DBY**	**UTY**	**EIF1AY**	**SMCY**	**DBY**	**UTY**	**EIF1AY**	**UBE1Y***
Total gene length [kb]	39.52	16.37	232.28	17.43	38.20	13.28	159.41	16.72	46.02	24.99	148.59	15.59	25.51
Number of target introns§	17	13	17	6	13	12	11	6	13	11	19	6	14
Number of microsatellites†	0	0	11	1	0	0	4	1	12	3	19	2	1
Number of YCATS¶	9	13	3	0	8	9	2	0	11	9	3	0	3
Total YCATS amplicon length [kb]¶	4.43	8.49	7.52	-	5.50	7.00	3.90	-	11.50**	4.56**	4.60**	-	0.82

§introns between 0.2 kb and 8 kb

†repeat unit 2-5 bp, >7 uninterrupted repeats, identified with the software Tandem Repeat Finder (Benson 1999)

¶Hellborg & Ellegren (2003)

*UBE1Y is not present in old world primates (Mitchell & Hammer 1996)

**Primer-BLAST in Genbank, NCBI

Summary of current methodologies

Despite a wide range of different methodological strategies employed to date, their success of developing male-specific markers has been limited. No predominate strategy seem to have emerged to date. Cytogenetic strategies have, despite their intriguing approach, rarely been employed. This is probably due to the fact that cytogenetic methods are time consuming and require special methodological know-how as well as laboratory facilities. Moreover, due to the multileveled approach they are particularly error-prone.

Only a few studies have screened genomic DNA libraries for male-specific markers, but most have failed to provide larger numbers of polymorphic Y-linked markers. In contrast, the more straightforward application of YCATS has been repeatedly applied and provided male-specificity. Yet, in general only short sequence data with little sequence diversity have been produced. As a result, YCATS have mainly been used in phylogeographic or even phylogenetic studies where a deeper evolutionary history among examined samples can be expected. The cross-species amplification strategy of known Y-markers has a high potential for researchers working on species with a genome sequencing project in a related taxon or other extensive Y sequence information. However, for most mammals this is currently not applicable and previous knowledge of the Y chromosome is missing.

Given the expected general interest of evolutionary biologists in disentangling paternal and maternal genetic lineages, the currently published work dealing with developing and employing Y-specific markers likely represents only a small portion of studies which have aimed to find male-specific polymorphism. There appears to be an obvious bias in that studies failing to develop such markers or showing no or very little variation are more difficult to get published. This publication bias, however, would be expected to be independent from methodological strategies.

2.4 Emerging methods

Current advances in molecular methods including next-generation sequencing (NGS) will enhance the discovery of Y-specific genetic markers. NGS refers to a group of alternative DNA sequencing technologies that are to the classical Sanger sequencing, and can generate hundreds of thousands of sequence reads at one time - thus increasing sequence capacity at an unprecedented rate (Hudson 2008; Shendure & Ji 2008). In this section, we describe unpublished and promising strategies for the development of Y-linked genetic markers in non-model organisms (bold pathways in Figure 2). Most of these strategies will benefit from NGS techniques by generating large amounts of Y sequence data and as such will increase the likelihood of discovering male-specific markers. Consequently, we focus in this section on the primary methodological step of obtaining longer DNA fragments for initial sequencing. Similar to the current methods, the emerging strategies can be divided into strategies screening simultaneously several individuals for sequence polymorphisms or strategies exploring large

sequence data from physically isolated material and libraries for microsatellites repeat motifs.

Beyond YCATS

Compared to the traditional YCATS approach (Hellborg & Ellegren 2003), longer stretches comprising several kb of Y-specific sequence data can be attained using long-range PCR to amplify single-copied MSY-specific genes (Wandeler and Camenisch, *in prep*.). This strategy fulfils instantly two of the three criteria for successful Y-marker development: male specificity and single-copy amplification, whilst the likelihood of finding sequence polymorphism among amplicons from different individuals is increased by obtaining longer sequence information. Additionally, intronic sequences may also contain polymorphic microsatellites (Luo *et al.* 2007; Wandeler & Camenisch in prep.). In Table 2, we present a list of single-copied MSY-linked genes as potential targets for this strategy. Publicly available Y chromosome reference sequences of MSY-linked genes from mouse, chimpanzee and human are used to estimate the proximate location of exonic and intronic sequences for one or several long-range PCR assays. Initial re-sequencing of short fragments of the selected gene-regions provides the necessary sequence information to design species- and Y-specific long-range PCR primers. Detailed sequence information is important for primer design, as in contrast to conventional PCR, long-range PCR requires perfectly matching primers. Re-sequencing can be done by applying newly-designed exonic primers or known conserved YCATS primers. Finally, male-specificity of long-range amplicons is verified and a few selected individuals are sequenced using Sanger sequencing. Alternatively, more amplicons from a large number of individuals can be pooled and sequenced simultaneously or sequenced using a parallel tagged sequencing approach on a NGS platform (Meyer *et al.*, 2008).

Y walking

We consider directional Y-chromosome walking as a promising strategy for generating large amounts of Y-specific sequence information. In this strategy, a known Y-linked sequence can be used as a starting point for sequencing into unknown flanking regions. Any DNA sequence tested for male-specificity including microsatellite flanking regions or sequences obtained by YCATS can serve as potential starting point. Again, by targeting regions near or within single-copied MSY-specific genes (Table 2), this strategy will likely provide Y-specific and single-copied amplicons. Several different genome-walking methodologies predominantly based on restriction digestion and PCR have been described, including inverse PCR (Triglia *et al.* 1988), ligation-mediated PCR (Rosenthal *et al.* 1990), and randomly primed PCR (Parker *et al.* 1991). Recently, these methods have been considerably improved (Reddy *et al.* 2008; Rampias *et al.* 2009; Tsuchiya *et al.* 2009). The sequence information gained by genome-walking methods is usually determined by the frequency of the cutting sites of the applied restriction enzymes. Since the number of Y-specific starting sequences can be a limiting factor, it is advantageous to maximize product size. However, as most methods are PCR based, product length barely exceeds 2 kb, although it should be possible to achieve considerably longer amplicons by

long-range PCR assays. Similar to the extended YCATS methods described above, amplicons can subsequently be sequenced by either conventional Sanger sequencing or by using a NGS platform.

Genomic Y sequence data

In species with no or very limited Y-sequence information, genomic Y data can be obtained by combining current methodological strategies with NGS technologies (Figure 2). The obvious advantage of these new technologies is the enormous increase in sequence output with lower costs and technical efforts. The sequencing of Y-chromosomal BAC or cosmid clones will be especially facilitated by high-throughput NGS. Moreover, even the whole Y could be decoded by *de novo* sequencing of a pool of hundreds of flow-sorted Y chromosomes, although it might not be possible to align the nucleotide reads to a single contig sequence given the highly repetitive structure of the Y chromosome (Skaletsky *et al.* 2003b). However, for the purpose of identifying microsatellite motifs this would be irrelevant provided that the selected NGS platform has the sufficient read length. Despite the high potential of these strategies for generating large amounts of Y sequence data, one main challenge remains. Although the obtained sequences are Y-chromosome derived, male-specificity and single-copy status of all sequences has to be verified before they are useful. This is labour-intensive as for example all microsatellite repeat motifs have to be tested for these criteria individually. Considering the architecture of the Y chromosome, the proportion of sequences fulfilling these criteria could be rather small. Nevertheless, these strategies represent a promising alternative to obtain Y-chromosomal data especially in species lacking any Y-sequence information so far.

2.5 Conclusions

Genetically tracing paternal lineages is hindered in most non-model species by the lack of Y chromosome markers despite the employment of a wide range of different methodological strategies. In the near future, the quest for Y-linked markers will benefit from a combination of recent technical advances such as NGS with current methods. For instance, longer Y-specific DNA sequence data by NGS can be obtained from long-range PCR products of intronic MSY genes, directional Y chromosome walking or from BAC libraries. Moreover, the amount of exonic and intronic Y sequence information as well as our knowledge of the Y chromosome architecture of different mammals will increase in the near future considering the growing number of genomes being sequenced.

Despite the promising potential of the presented current and emerging methodological strategies, there is likely no straightforward solution in obtaining Y-linked genetic markers. The distinct architecture of the Y chromosome with its highly palindromic structure, the widespread sequence homologies to the X chromosome and the general low levels of genetic variation observed will hamper the discovery of genetic markers fulfilling our criteria for Y-

linked loci. Furthermore, the evolutionary history of the study species will further affect the observed level of Y chromosome variation. However, although discovering Y-linked genetic markers is difficult, the efforts are worthwhile considering their apparent potential to explore sex-biased dispersal patterns and independent demographic population histories of males and females in wild animal populations.

In comparative mythology (Campell 1949), a quest describes a heroes' journey in which "A hero ventures forth from the world of common day into a region of supernatural wonder: fabulous forces are there encountered and a decisive victory is won: the hero comes back from this mysterious adventure with the power to bestow boons on his fellow man." We advise researchers with limited resources embarking on such a quest for the Y to form collaborations with laboratories in which the techniques presented in this review are well established. Additionally, the mating system and life history of the species in *quest*ion also needs to be carefully considered upon embarkation. Following these two rules should enable researchers to bestow a wealth of suitable Y-linked markers on their fellow scientists.

Acknowledgements

The authors are grateful to Glauco Camenisch (Zoological Museum, Zurich), Angelika Schwarze and Beat Steinmann (Children's Hospital, Zurich), Patricia O'Brien and Malcom Ferguson-Smith (Veterinary School, Cambridge) for their support. We appreciated the valuable comments made on the manuscript by Briana Gross, Anna Lindholm, and four anonymous reviewers. This work was supported by a Basler Stiftung für Biologische Forschung grant to PW and A.H. Schultz Foundation grants to MK and MG.

Author Contributions

MPG, MK, and PW conceived the study. MPG and PW wrote the manuscript. MK edited the manuscript. MPG compiled tables and figures. CS and APS provided analytical inputs.

Chapter 3

Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms

Maja P. Greminger[1], Kai N. Stölting[2], Alexander Nater[1], Benoit Goossens[3,4,5], Natasha Arora[1], Rémy Bruggmann[6,7], Andrea Patrignani[6], Beatrice Nussberger[8], Reeta Sharma[9], Robert H. S. Kraus[10], Laurentius N. Ambu[5], Ian Singleton[11,12], Lounes Chikhi[9,13,14], Carel P. van Schaik[1] and Michael Krützen[1]

[1]Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

[2]Unit of Ecology & Evolution, Department of Biology, University of Fribourg, Fribourg, Switzerland

[3]Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom

[4]Danau Girang Field Centre, c/o Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[5]Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[6]Functional Genomics Center, University of Zurich, Zurich, Switzerland

[7]Department of Biology, University of Berne, Berne, Switzerland

[8]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[9]Population and Conservation Genetics, Instituto Gulbenkian de Ciencia, Oeiras, Portugal

[10]Conservation Genetics Group, Senckenberg Research Institute and Natural History Museum, Gelnhausen, Germany

[11]Foundation for a Sustainable Ecosystem (YEL), Medan, Indonesia

[12]PanEco, Foundation for Sustainable Development and Intercultural Exchange, Berg am Irchel, Switzerland

[13]CNRS, Laboratoire Evolution and Diversité Biologique, Université Paul Sabatier, Toulouse, France

[14]Université de Toulouse, Toulouse, France

3.1 Abstract

High-throughput sequencing has opened up exciting possibilities in population and conservation genetics by enabling the assessment of genetic variation at genome-wide scales. One approach to reduce genome complexity, i.e. investigating only parts of the genome, is reduced-representation library (RRL) sequencing. Like similar approaches, RRL sequencing reduces ascertainment bias due to simultaneous discovery and genotyping of single-nucleotide polymorphisms (SNPs) and does not require reference genomes. Yet, generating such datasets remains challenging due to laboratory and bioinformatical issues. In the laboratory, current protocols require improvements with regards to sequencing homologous fragments to reduce the number of missing genotypes. From the bioinformatical perspective, the reliance of most studies on a single SNP caller disregards the possibility that different algorithms may produce disparate SNP datasets. We present an improved RRL (iRRL) protocol that maximizes the generation of homologous DNA sequences, thus achieving improved genotyping-by-sequencing efficiency. Our modifications facilitate generation of single-sample libraries, enabling individual genotype assignments instead of pooled-sample analysis. We sequenced ~1% of the orangutan genome with 41-fold median coverage in 31 wild-born individuals from two populations. SNPs and genotypes were called using three different algorithms. We obtained substantially different SNP datasets depending on the SNP caller. Genotype validations revealed that the *Unified Genotyper* of the *Genome Analysis Toolkit* and *SAMtools* performed significantly better than a caller from *CLC Genomics Workbench* (CLC). Of all conflicting genotype calls, CLC was only correct in 17% of the cases. Furthermore, conflicting genotypes between two algorithms showed a systematic bias in that one caller almost exclusively assigned heterozygotes, while the other one almost exclusively assigned homozygotes. Our enhanced iRRL approach greatly facilitates genotyping-by-sequencing and thus direct estimates of allele frequencies. Our direct comparison of three commonly used SNP callers emphasizes the need to question the accuracy of SNP and genotype calling, as we obtained considerably different SNP datasets depending on caller algorithms, sequencing depths and filtering criteria. These differences affected scans for signatures of natural selection, but will also exert undue influences on demographic inferences. This study presents the first effort to generate a population genomic dataset for wild-born orangutans with known population provenance.

3.2 Introduction

The availability of high-throughput sequencing has revolutionized the fields of population genetics and molecular ecology (Seeb *et al.* 2011a). Early genomic work focused mainly on broad comparative analyses between species (e.g. Nielsen *et al.* 2005a; Bakewell *et al.* 2007; Gibbs *et al.* 2007; Kosiol *et al.* 2008; Enard *et al.* 2010) and was limited to one or a few individuals per species. The emergent field of population genomics (Hohenlohe *et al.* 2010), including conservation (Steiner *et al.* 2013) and landscape genomics (Schoville *et al.* 2012), investigates genomic allele-frequency patterns at the species level, i.e. among and within natural populations. Main interests revolve around exploring patterns of genetic diversity, differentiation and admixture, inferring demographic population histories, and studying signals of local adaptations in wild populations (Hohenlohe *et al.* 2010; Schoville *et al.* 2012; Steiner *et al.* 2013).

To date, population-genomics studies have mainly focused on humans (Altshuler *et al.* 2010a; Altshuler *et al.* 2010b), some model species (e.g. Weigel & Mott 2009; Cao *et al.* 2011; Kumar *et al.* 2012) and others relevant to agricultural production (e.g. Gibbs *et al.* 2009; Elferink *et al.* 2012). Other taxa, particularly those with large genomes, have remained largely unexplored because of significant challenges in the laboratory and during bioinformatical analyses (Helyar *et al.* 2011; Nielsen *et al.* 2011; Steiner *et al.* 2013). Sequencing of complete genomes of many individuals is usually still prohibitive because of associated costs and bioinformatical complexities, especially in species where a reference genome is unavailable. Yet, many biological questions can be addressed by describing polymorphisms from a subset of genomic regions, provided that these regions are approximately evenly distributed throughout the genome.

In the laboratory, several strategies have recently been developed enabling so-called 'reduced genome complexity sequencing', i.e. sampling only a small fraction of the genome in several individuals. These strategies include sequencing of reduced-representation libraries (RRLs) (van Tassell *et al.* 2008), restriction-site-associated DNA sequencing (Baird *et al.* 2008; Hohenlohe *et al.* 2010), and other sequence-based-genotyping approaches (Elshire *et al.* 2011; Truong *et al.* 2012). Essentially, all of these methods are based on the same key principle: reducing genome complexity by digestion of genomic DNA with one or several restriction enzymes followed by a selection of resulting restriction fragments, and high-throughput sequencing of the final set of fragments.

One of the key characteristics of the aforementioned methods is that, at least in theory, read mapping can be carried out regardless of the availability of a reference genome by constructing a reference sequence from overlapping sequence stacks (e.g. van Bers *et al.* 2010; Young *et al.* 2010; Kerstens *et al.* 2011; Truong *et al.* 2012; Senn *et al.* 2013). Moreover, the similarity among sequence stacks of different individuals allows the direct estimation of allele frequencies by simultaneous identification of polymorphisms and genotype calling (genotyping-by-sequencing). This reduces the major issue of ascertainment bias, which arises

when markers are identified in a small subset of individuals and subsequently genotyped in an extended sample set (Helyar *et al.* 2011; Seeb *et al.* 2011a; Seeb *et al.* 2011b).

One popular reduced-genome complexity approach is RRL sequencing. RRLs were first used to generate single-nucleotide polymorphisms (SNP) maps of the human genome using classical Sanger sequencing (Altshuler *et al.* 2000). Since Van Tassel *et al.* (van Tassell *et al.* 2008) first adapted the approach to high-throughput sequencing, it has been applied in a number of SNP discovery studies (e.g. Amaral *et al.* 2009; Kerstens *et al.* 2009; Hyten *et al.* 2010; van Bers *et al.* 2010; Kraus *et al.* 2011). In the RRL approach, the number of restriction fragments subjected to high-throughput sequencing is reduced via size-selection before sequencing library preparation. RRLs allow the degree of complexity reduction to be customized by defining the selected fragment-size range. By providing easy access to flanking sequences necessary to design SNP genotyping assays when a reference genome is unavailable, RRLs are superior to other reduced-complexity approaches (Baird *et al.* 2008; Hohenlohe *et al.* 2010). In the RRL approach, long DNA stretches can be sequenced by simply size-selecting for longer fragments (up to several kb possible) and complete sequencing of these fragments independent of the platform read length through shearing of fragments prior to high-throughput sequencing library preparation followed by assembly of the resulting sequence fragments (Kerstens *et al.* 2009).

Although the RRL principle is highly promising for generating population genomic SNP data, current protocols must be improved so as to i) facilitate library construction for individual samples, and most importantly, ii) maximize the number of homologous fragments generated during library construction. In the past, RRL sequencing has usually been performed on pools of DNA samples from multiple individuals for practical reasons (e.g. Wiedmann *et al.* 2008; van Bers *et al.* 2010; Kerstens *et al.* 2011; Kraus *et al.* 2011). However, pooling leads to the loss of major biological information as it prohibits the assignment of individual genotypes (i.e. genotyping-by-sequencing). Because of this, many biological questions, such as investigating admixture or linking phenotypes with genotypes in studies of natural selection, cannot be addressed when samples are pooled. Furthermore, pooling strongly increases the risk of missing rare alleles, especially if there are many individuals in the pool (Cutler & Jensen 2010). In addition, pooling is highly sensitive to variation in DNA concentration among samples, which will inadvertently lead to an over – or underrepresentation of certain alleles (Cutler & Jensen 2010). Thus, current protocols need to be improved to facilitate RRL generation of individual samples.

Analyzing individual samples requires improvements to minimize DNA loss during purification steps, which is particularly important if sample-DNA quantity is limited. Moreover, genome complexity needs to be reduced in a reproducible manner (i.e. homologous sites must be sequenced) across samples as this primarily determines the effectiveness of the genotyping-by-sequencing principle and reference-free mapping (Elshire *et al.* 2011). Non-overlapping sequences will lead to a high number of missing genotypes. The accurate sequencing of homologous sites is also of particular importance when working with pooled samples, as the

true number of sequenced individuals at a particular SNP site cannot be determined. In the most extreme case, only alleles of one individual would be sequenced. In such a case, however, allele frequencies would nonetheless be estimated under the assumption that all allele copies in the pool had been sampled.

From a bioinformatical point of view, the amounts of raw data produced by high-throughput sequencing platforms are vast and many computational steps are required to translate raw outputs into high-quality SNP calls (Nielsen *et al.* 2011). Thus, accurately identifying SNPs and calling genotypes from high-throughput sequencing data while filtering out sequencing errors remains a challenge. Various SNP calling programs have been introduced and algorithms are under constant development (Li *et al.* 2009; DePristo *et al.* 2011; Nielsen *et al.* 2011; Pabinger *et al.* 2013).

One of the most widely used commercial software suites for genomic data analysis is the *CLC Genomics Workbench* (CLC bio, Aarhus, Denmark). The software contains a basic SNP caller (hereafter referred to as 'CLC') that detects SNPs based solely on applying quality thresholds to sequencing, mapping and base quality. Genotypes are determined using hard-filter criteria, i.e. by simply counting the number of sequencing reads for each allele and applying arbitrary custom cut-off rules. For instance, a genotype would be called heterozygous if an alternative allele is present in 20-80% of the reads. However, for low sequencing depths this way of genotype calling tends to underestimate the number of heterozygous genotypes (Nielsen *et al.* 2011).

Arguably, two of the most popular non-commercial software suites are the *Genome Analysis Toolkit* (Broad Institute) (McKenna *et al.* 2010; DePristo *et al.* 2011) and *SAMtools* (Li *et al.* 2009). Both *SAMtools* and the *Unified Genotyper* of the *Genome Analysis Toolkit* (hereafter referred to as 'GATK'), incorporate uncertainty in a probabilistic framework, in order to call SNPs and genotypes simultaneously (Li *et al.* 2009; McKenna *et al.* 2010; DePristo *et al.* 2011). Both SAMtools and GATK allow the joint analysis of all samples from one population (multi-sample calling). A major strength of the Bayesian framework is the potential to incorporate prior information, such as previous observations of alternative alleles, heterozygosity, and allele frequencies. Ideally, additional information such as representative reference SNPs or linkage-disequilibrium patterns could be incorporated (DePristo *et al.* 2011; Le & Durbin 2011; Nielsen *et al.* 2011). Unfortunately, such information so far limited to a few model species (e.g. Arabidopsis (Weigel & Mott 2009)) and humans (Altshuler *et al.* 2010a). It has been proposed that in contrast to CLC, GATK (and potentially also SAMtools) might have the tendency to overestimate the number of heterozygous genotypes (DePristo *et al.* 2011). This is because GATK aggressively calls alternative alleles in favor of high sensitivity, resulting in a high number of false-positive calls which require extensive post-filtering.

Despite the fact that accurate SNP and genotype calling is fundamental for precise population parameter estimation in downstream analyses (Nielsen 2005; Pool *et al.* 2010; Helyar *et al.* 2011), to our knowledge direct comparisons of different SNP callers in the aforementioned

context are still scarce. To date, most studies employ only one SNP caller, although it is conceivable that different callers will produce different datasets. In previous studies, validations were often restricted to confirming and comparing the polymorphic state of SNPs (e.g. Sanchez *et al.* 2009; Hyten *et al.* 2010; van Bers *et al.* 2010), but not actual genotypes at the individual level.

Here, we provide a comprehensive framework to obtain high-quality SNP data in population genomics, addressing both laboratory and bioinformatic challenges. First, we refined and improved an RRL protocol (iRRL), which maximizes the generation of homologous DNA fragments across individuals, thus achieving high genotyping-by-sequencing efficiency. Our protocol also contains modifications for economical handling of DNA during library preparation. All modifications support the establishment of single-sample libraries. Second, we directly compared three popular SNP callers (GATK, SAMtools and CLC) using our iRRL data generated for two orangutan populations (Genus: *Pongo*).

Orangutans are the only great apes found outside Africa and the phylogenetically most distant great apes to humans, which makes them particularly interesting to study in terms of the evolution of the hominid lineage (Delgado & van Schaik 2000; Groves 2001). In contrast to humans (e.g. the *International HapMap Project* (Altshuler *et al.* 2010a); the *1000 Genomes Project* (Altshuler *et al.* 2010b)), in non-human great apes large-scale population genomic data from wild-born individuals with known population origin are scarce (but see Hvilsom *et al.* 2012; Prado-Martinez *et al.* 2013). Rather, most genomic data were generated from a small number of zoo animals with mostly unknown population origins (Locke *et al.* 2011; Auton *et al.* 2012; Prufer *et al.* 2012; Scally *et al.* 2012), thus providing a limited perspective for population genomic analyses of wild populations. Genome-wide data in orangutans will enable the investigation of the genetic basis of local adaptations among orangutan populations (van Schaik *et al.* 2009b). Moreover, population genomic data will shed more light on the particularly complex demographic history of orangutans, as shaped by volcanic eruptions and recurrent sea level changes connecting the islands of Borneo and Sumatra during the Pleistocene (Steiper 2006; Arora *et al.* 2010; Nater *et al.* 2011; Nater 2012; Nater *et al.* 2013).

3.3 Results

Improved reduced-representation sequencing

We developed a protocol to construct improved RRLs (referred to as iRRLs) that maximizes efficiency and repeatability of genome complexity reduction. We applied several key modifications to the method outlined in van Tassel *et al.* (van Tassell *et al.* 2008) including: (i) high-resolution fragment-size-selection down to an accuracy of one base pair to increase precision of isolating homologous fragments (Figure S2), (ii) modifications to minimize DNA loss during purification steps, achieving DNA recovery rates of >95%, and (iii) adjustments to

establish single-sample libraries to avoid the necessity of sample pooling. In order for restriction enzymes to generate homologous fragments across samples, our protocol includes recommendations for suitable sample handling and DNA isolation to avoid DNA strand breaking prior to digestion.

We established iRRLs for 31 unrelated orangutans from two populations, the West Alas population on northwestern Sumatra (WA, *Pongo abelii*, n=15) and the South Kinabatangan population on northeastern Borneo (SK, *Pongo pygmaeus*, n=16; Figure 1, Table S1). Based on the number of study individuals, the orangutan genome size of 3.09 Gigabases (Gb) (Locke *et al.* 2011), the budgeted SOLiD4 sequencing costs, and an intended 30-fold (30x) sequencing depth, we calculated our targeted degree of genome complexity reduction to be 100-fold, i.e. 1% of the genome. We carried out *in-silico* digests of the orangutan reference genome (Locke *et al.* 2011) with several candidate blunt-end cutters in order to identify the restriction enzyme suitable to our project needs (see Methods). In the selected size range of 104-123 bp, a *Hae*III digest yielded 305,574 predicted fragments with low repetitive sequence content (representing the desired 1.07% of the genome, Figure 2). Our *in-silico* digest demonstrated the importance of uniform fragment selection. For instance, extending the selected size range by as few as 4 bp (e.g. 100-123 bp) in all individuals would have already resulted in a 25% increase in the selected genome proportion, i.e. 1.32% of the genome with lower average coverage per site. Furthermore, a range shift of a few base pairs in some individuals in either direction would lead to a dramatic decrease in homology among the generated fragments.

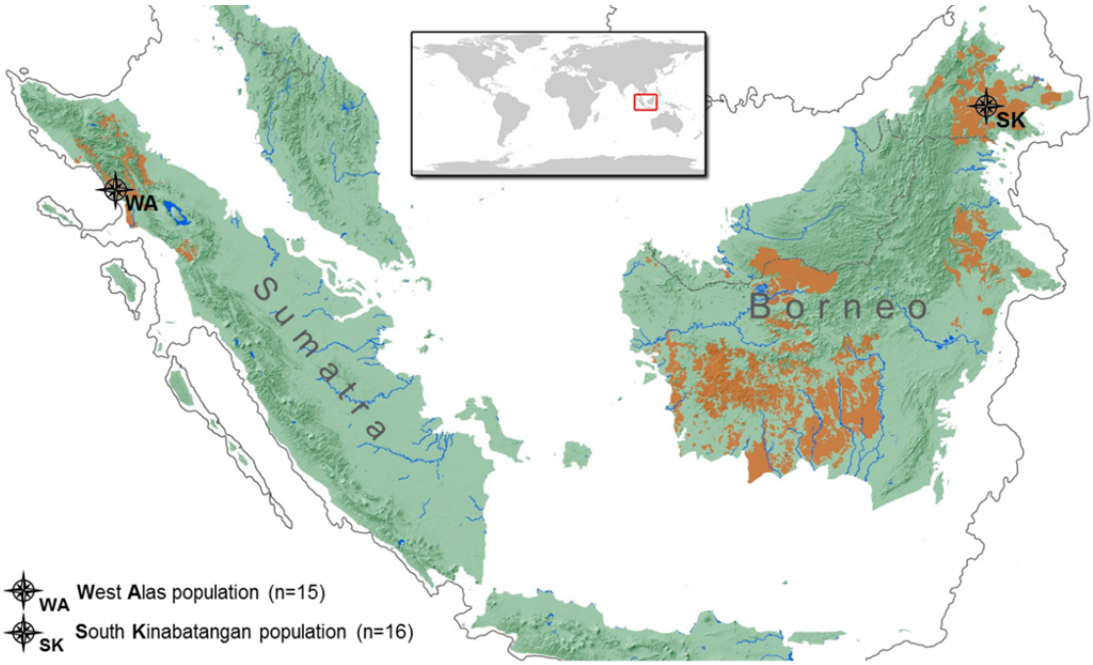


Figure 1. Geographic location of the two orangutan study populations. The areas colored in brown indicate the current distribution of orangutans.

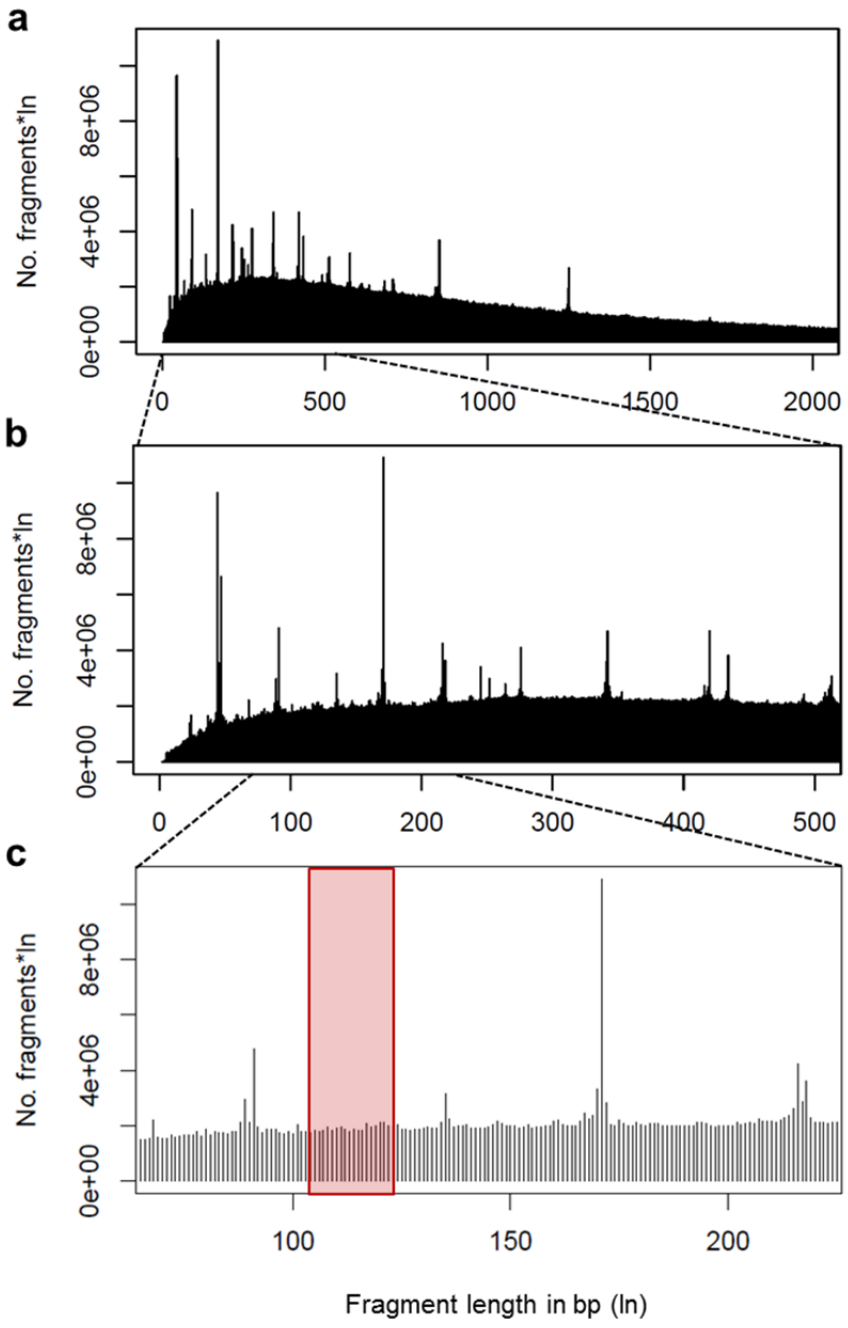


Figure 2. *In-silico Hae*III digest of the orangutan reference genome. Panel a, b and c represent increasing levels of details. The x- and y axis show the generated fragment lengths in base pairs and the number of fragments multiplied by fragment length, respectively. Peaks are due to repetitive sequences. The isolated fragment size range (104-123 bp) is indicated in red.

In total, we obtained 675 million beads for the West Alas population and 762 million beads for the South Kinabatangan population by individually barcoding iRRLs and sequencing them on the SOLiD4 platform (Life Technologies) with paired-end chemistry. Raw sequence data were submitted to the NCBI Sequence Read Archive [BioProject: PRJNA230877; BioSamples: SAMN02439270-SAMN02439300]. Median numbers of mapped reads for each individual

were 32,345,177 for the West Alas population and 43,451,986 for the South Kinabatangan population (Tables 1 and S2). The greater sequencing output for South Kinabatangan individuals is related to different performances of our SOLiD4 runs that were beyond our control. We also observed a poor performance of the F5 sequence read direction. We only considered high quality base pairs (bp_{hiqual}) in downstream analyses, i.e. sites with mapping and base quality phred scores of ≥ 30, and a minimal sequence depth of 10x. Applying these stringent filters, we retained 10,930,563 bp_{hiqual} with 41x median sequence coverage for West Alas individuals and 18,186,855 bp_{hiqual} with 42x median coverage for South Kinabatangan individuals (Table S2).

To assess the performance of our iRRL protocol, we estimated the iRRL target efficiency as the percentage of obtained bp_{hiqual} sites which were predicted by the *in-silico* digest (= target sites). iRRL efficiency varied among individuals but was very high with a median of 97% for West Alas individuals and 86% for the South Kinabatangan individuals (Tables 1 and S2). Thus, the vast majority of sequenced high quality bases were target sites, i.e. predicted by the *in-silico* digest of the orangutan reference genome.

Table 1. Overview of the sequencing of improved reduced-representation libraries (iRRLs) for the West Alas (WA) and South Kinabatangan (SK) orangutan study populations.

	Pop_WA (Sumatra)	Pop_SK (Borneo)
No. of individuals	15	16
iRRL stacks per individual (predicted)[a]	305,574	305,574
Median iRRL target efficiency[b]	97%	86%
Total no. of beads per population	675,295,801	762,234,081
Total no. of mapped reads per population	528,081,935	646,922,204
Median no. of mapped reads per individual	32,345,177	43,451,986
% reads mapped F3/F5 (mappability)[c]	74.9/7.3	67.0/17.0
Mean no. of bp_{hiqual} per individual[d]	10,930,563	18,186,855
Median sequence coverage per individual[e]	41x	42x

[a]Predicted by *in-silico* digest of the orangutan reference genome *ponabe2* (Sumatran) with *Hae*III
[b]Percentage of sequenced sites that were predicted by the *in-silico* digest
[c]F3/F5 are the sequence read directions of the paired end sequencing mode
[d]Number of sequenced base pairs passing all high quality filters (sites used for SNP detection)
[e] GATK estimates based on bp_{hiqual}

Comparison of SNP discovery and genotype calling

We identified SNPs *de-novo* and called individual genotypes using three different algorithms: GATK, SAMtools, and CLC. Calls were based on the stringent bp_{hiqual} filter thresholds. For the GATK and SAMtools dataset, we also applied a minimal threshold on the genotype quality score (GQ ≥ 10). In addition, we performed identical population-based filtering for all three algorithms. We only accepted SNPs with a maximum of two alleles and genotypes meeting all

quality filter criteria in at least eight individuals per population (n ≥ 16 chromosomes), allowing accurate allele frequency estimations. Applying all filters we retrieved 57,396 SNPs in the GATK dataset, 75,364 SNPs in the CLC dataset, and 24,103 SNPs in the SAMtools dataset (Table 2).

Compared to similar studies (e.g. Wiedmann *et al.* 2008; Kerstens *et al.* 2009; Sanchez *et al.* 2009; Esteve-Codina *et al.* 2011), median sequence coverage at SNP sites across all individuals in our datasets was extremely high (82x for GATK, 48x for CLC, and 27x for SAMtools), although coverage counts differed drastically among datasets. This discrepancy in coverage counts could be attributed to a different treatment of quality scores in read counting and/or different default parameters among the callers, since we applied identical quality thresholds to the data. The considerably lower read counts in the SAMtools dataset and potentially different prior probabilities in the Bayesian framework may be causal for the strikingly lower number of total SNPs in our SAMtools dataset.

We observed a low overlap of SNPs among the three datasets, i.e. SNP sites present in at least two datasets irrespective of the genotype calls at the individual level (Figure 3). In total, 18,482 SNPs overlapped among all three datasets. At only 13%, the SAMtools dataset exhibited the lowest percentage of private SNPs compared to the other two algorithms (Figure 3).

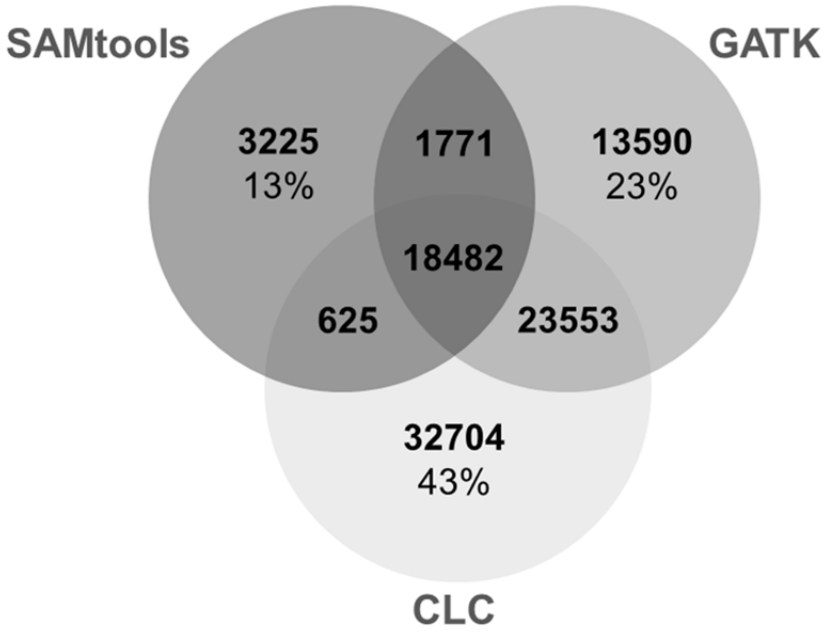


Figure 3. Overlap of SNPs among the datasets obtained from three different callers. Percentages specify the proportion of SNPs exclusively present in the particular dataset for each caller.

Table 2. Overview of SNP discovery and genotype calling using three different callers. We required all SNPs to have a genotype call passing all stringent quality filters in a minimum of eight individuals per population (population-based filtering). The intersect datasets contain exclusively concordant genotype calls between the designated SNP callers. Pop_SK: South Kinabatangan population, Pop_WA: West Alas population.

	GATK_v.2.5-0			CLC_v.5.0.1			SAMtools_v.0.1.19		
	Pop_SK	**Pop_WA**	**Overall**	**Pop_SK**	**Pop_WA**	**Overall**	**Pop_SK**	**Pop_WA**	**Overall**
No. of SNPs	34257	40248	57396	34788	55585	75364	14494	14903	24103
No. of private SNPs	17148	23139	40287	19779	40576	60355	9200	9609	18809
% singletons	7.68	10.83	12.18	11.53	27.47	25.59	14.63	21.66	22.19
Median site heterozygosity[a]	0.267	0.250	/	0.236	0.200	/	0.266	0.231	/
Median coverage per individual	93x	70x	82x	66x	29x	48x	66x	19x	27x

	GATK-CLC$_{intersect}$			SAMtools- GATK$_{intersect}$			SAMtools-CLC$_{intersect}$		
	Pop_SK	**Pop_WA**	**Overall**	**Pop_SK**	**Pop_WA**	**Overall**	**Pop_SK**	**Pop_WA**	**Overall**
No. of SNPs	21475	24936	37085	11325	12350	18933	9861	11310	17163
No. of private SNPs	12149	15610	27759	6583	7608	14191	5853	7302	13155
% singletons	9.91	17.98	12.82	9.99	20.53	19.37	10.54	23.08	21.60
Median site heterozygosity[a]	0.250	0.222	/	0.286	0.231	/	0.266	0.222	/
Median coverage per individual[b]	107x (65)	81x (27)	96x (37)	55x (98)	18x (98)	20x (99)	69x (76)	19x (35)	26x (46)

[a]Based on the sites being polymorphic within the population

[b]Coverage values of intersect datasets are taken from the first named SNP caller. The coverage values of the second named caller are given in brackets

Many of the non-overlapping sites were present in initial SNP discoveries, but were removed because less than eight individuals per population had a genotype call meeting all high-quality filter criteria (population-based filter). For the CLC and SAMtools dataset, genotype calls often failed the minimum coverage requirement of 10 reads. For the GATK dataset, many genotype calls did not have a sufficiently high genotype quality score.

For all overlapping SNPs, we evaluated the concordance of genotype assignments by comparing for each individual whether two callers produced identical genotypes. The percentage of identical genotype calls varied among individuals with median values of 97.51% for GATK-CLC, 98.32% for SAMtools-GATK, and 97.24% for SAMtools-CLC (Tables 3 and S3-5). A quantitative investigation of discordantly called genotypes between the callers revealed that the vast majority (> 99.77%) of these genotypes were called heterozygous by one caller but homozygous for either of the alleles by the other caller. The relative distribution of these heterozygous/homozygous genotype calls appeared to be strongly biased (Figure 4). For example, examining discordant genotype calls between GATK and CLC showed that in most cases (93.02%), GATK assigned a heterozygous genotype while CLC assigned a homozygous one. Pairwise SNP caller comparisons revealed that SAMtools had the highest tendency to call heterozygotes in such cases, followed by GATK and CLC (Figure 4).

Table 3. Median genotype concordance between designated SNP callers for overlapping SNP sites assessed at the individual level.

	GATK-CLC	SAMtools- GATK	SAMtools-CLC
% same genotype called Pop_WA	96.92	98.46	96.15
% same genotype called Pop_SK	98.27	98.04	97.45
% same genotype called overall	97.51	98.32	97.24
% same genotype called overall (range)	93.59-98.38	97.08-99.26	92.46-97.82

We also created three intersect datasets by accepting only identically assigned genotypes between pairs of SNP callers (at the individual level prior to the population-based filtering). This procedure has been suggested to reduce caller-specific errors and increase specificity (e.g. Nielsen *et al.* 2011). We retained 37,085 SNPs for the GATK-CLC$_{intersect}$ dataset, 18,933 SNPs for the SAMtools-GATK$_{intersect}$ dataset, and 17,163 SNPs for the SAMtools-CLC$_{intersect}$ dataset (Table 2).

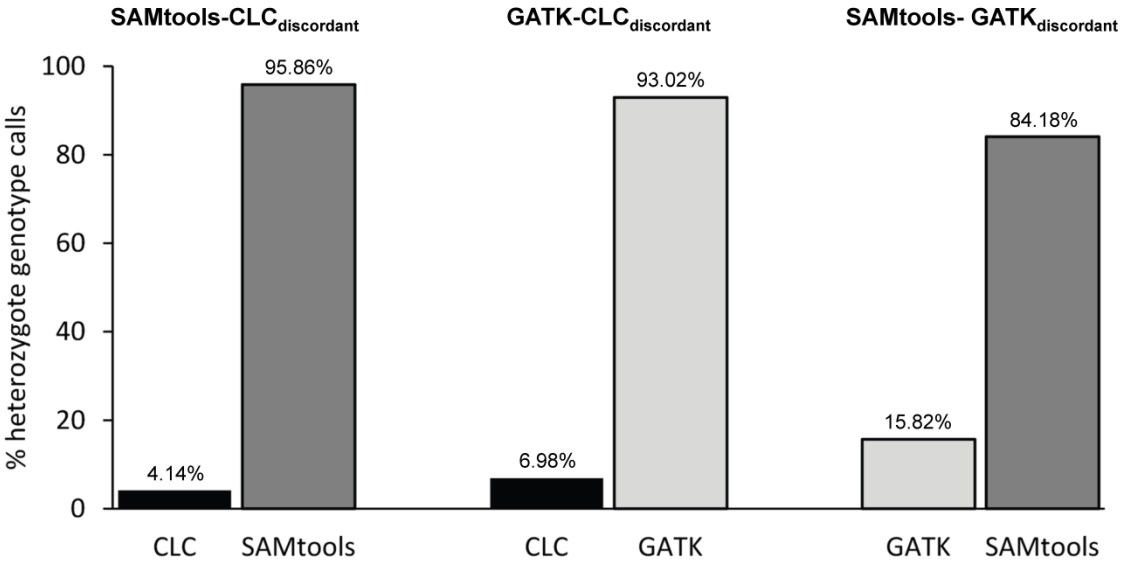


Figure 4. Quantitative investigation of discordant genotype calls between pairs of SNP callers. For the vast majority (> 99.77%) of discordant genotype calls, one caller assigned a heterozygous genotype but the other caller a homozygous genotype for either of the alleles. The y-axis represents the percentage of heterozygous genotype calls in such cases. The values are median numbers across all study individuals.

Impact on biological inferences

Over all six datasets, there were more sites segregating in the Sumatran West Alas population compared to the Bornean South Kinabatangan population (Table 2). The vast majority of SNPs (70-80% depending on the dataset) were private. In addition, we observed a large percentage of singletons (Table 2). The highest number of singletons was obtained in the CLC dataset (26%) followed by SAMtools (22%) and GATK (12%). Median site heterozygosity was always higher for the South Kinabatangan population than for the West Alas population.

To investigate the potential impact of the different SNP datasets on biological downstream analyses, we calculated three important statistics. (i) Kernel-density distributions for site heterozygosity and (ii) minor allele frequency were not identical among the SNP datasets (Permutation test of equality, $p<<0.001$, Figure 5). From a qualitative point of view, differences in kernel density distributions among all six datasets were especially pronounced for the West Alas population (Figure 5a,c) for which median sequence coverage was lower compared to the South Kinabatangan population. Nevertheless, it is striking that we obtained these differences despite a stringent minimal read cut-off of 10 reads and 29x (CLC value) medium sequence coverage. For example, the CLC dataset consisted of the largest proportion of low frequency alleles. In contrast, GATK called more variants at mid-frequency and showed higher overall heterozygosity levels.

To evaluate the impact of the SNP dataset differences on genome-wide scans for signatures of natural selection, we performed (iii) sliding-window analyses (100 kb windows, 25 kb step size) to identify signals of putative selective sweeps based on population differentiation. We used the allele-frequency differential (*D)* to measure population differentiation. We arbitrarily defined outlier regions as windows with an average population differentiation $D > 0.95$ (covered by at least 2 SNPs). The overlap of outlier windows among datasets was low. Only 3.8% of all detected outlier windows were identical among all three single-caller datasets (Figure 6), which improved to 13.5% when intersect datasets were used (Figure S3).

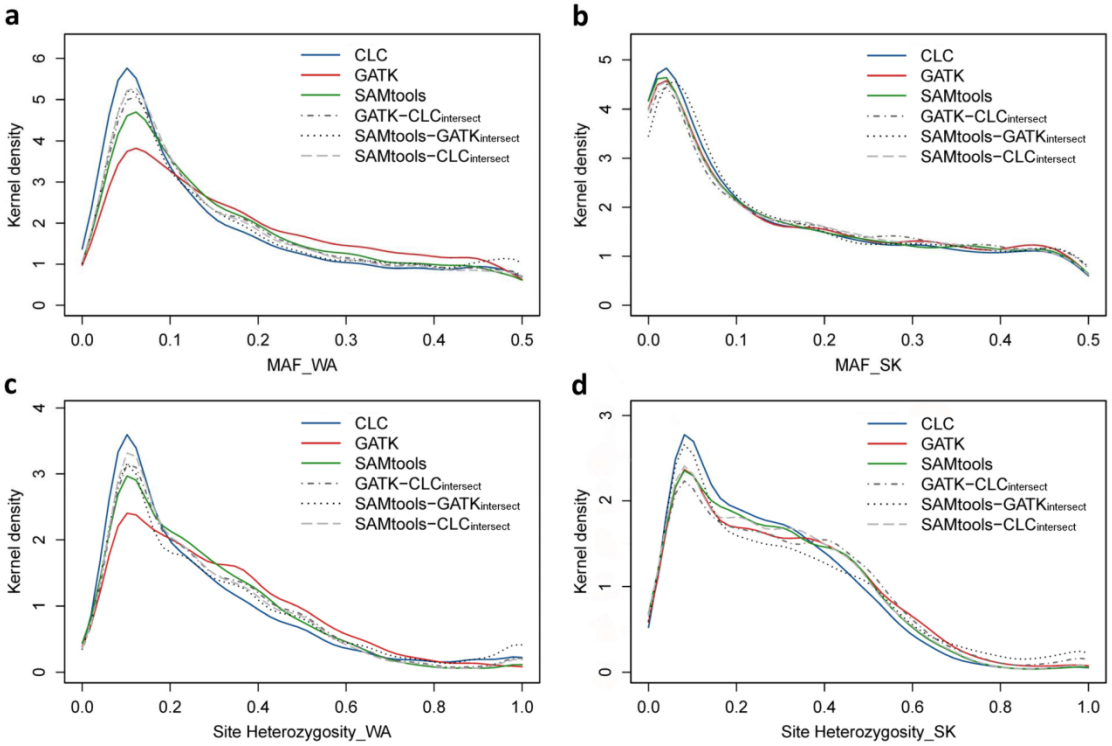


Figure 5. Kernel density distributions of minor-allele frequency and site heterozygosity using the different SNP datasets. For each of the six SNP data sets (CLC, GATK, SAMtools, GATK-CLCintersect, SAMtools-GATKintersect, and SAMtools-CLCintersect) we computed the minor-allele frequency (MAF) for the Sumatran (WA) and Bornean (SK) individuals (panels a and b, respectively), and site heterozygosity for WA and SK (panels c and d, respectively).

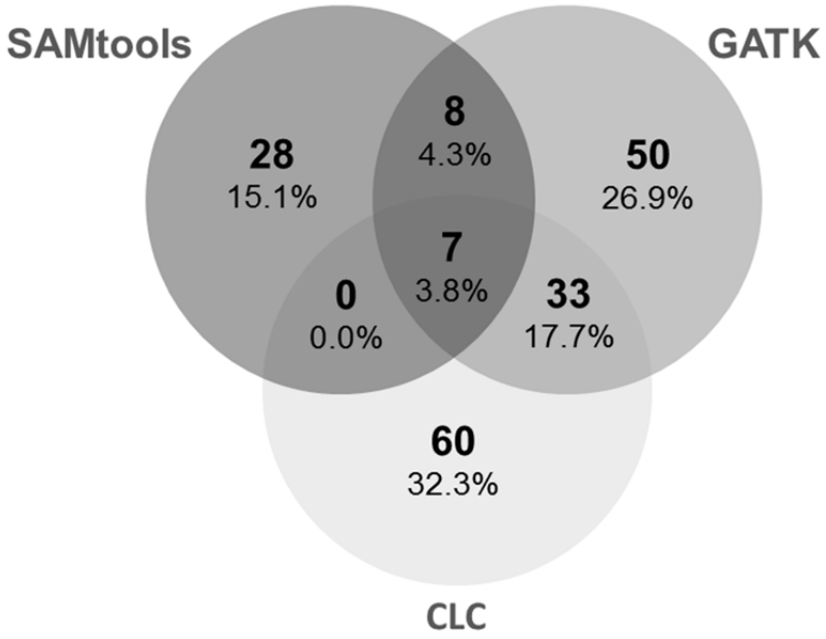


Figure 6. Overlap of outlier regions among SNP datasets in genome-wide scans for positive selection. For all SNP datasets we performed sliding-window analyses (100 kb window, 25 kb step size) of the absolute allele-frequency differential (D) between the SK and WA population. All windows with an average window D > 0.95 were considered as outliers, i.e. candidate regions for selective sweeps. Percentage values are given in relation to the total number of outlier windows.

Genotype validations

To determine genotype accuracy, we validated 63 genotypes from a subset of 58 SNPs overlapping among datasets by classical Sanger sequencing. We picked SNPs with the only requirements that a minimum of ten individuals per population had an assigned genotype and that at least one individual showed a conflicting genotype call between GATK/SAMtools and CLC. Because all validated genotypes were identical between GATK and SAMtools, we did not distinguish between the two for this analysis, but rather focused on the difference between probabilistic (GATK and SAMtools) and hard-filtering (CLC) callers.

Our results show that GATK/SAMtools clearly outperformed CLC, with a correct genotype assignment in 83% of the conflicting calls (Table 4). GATK/SAMtools calling accuracy was especially high for singletons (92% true in GATK/SAMtools, 8% true in CLC) and for genotypes that were according to GATK/SAMtools homozygous for either of the two alleles but heterozygous according to CLC (89% true in GATK/SAMtools, 11% true in CLC). We also verified the genotype accuracy of identical calls and found 4 miscalled genotypes out of 114 (3.5%).

Table 4. Overview of genotype validations at overlapping SNP sites.

Category	SNPs validated	Genotypes validated	True CLC		True GATK/SAMtools	
			n	%	n	%
Discordant calls[a]						
Singleton site determined by GATK/SAMtools [b]	8	8	1	12.5	7	87.5
Singleton site determined by CLC[b]	4	4	0	0	4	100
Homozygote with GATK/SAMtools but heterozygote with CLC	23	28	3	10.71	25	89.29
Heterozygote with GATK/SAMtools but homozygote with CLC	23	23	7	30.43	16	69.57
Total	**58**	**63**	**11**	**17.46**	**52**	**82.54**
Concordant calls[c]						
Total	**53**	**114**		**110 (96.49%)**		

[a]Overlapping SNP sites but discordant genotype assignments

[b]Loci were exclusively counted in this category without considering them in the homo- or heterozygote categories below

[c]100 of the 114 genotypes were validated from the same sites used to validate the discordant genotypes. The remaining 14 genotypes were validated from 14 SNPs chosen randomly from the GATK-$CLC_{intersect}$ dataset (exclusively identical genotype calls)

Characteristics of SNP callers

GATK seemed to be conservative in calling singletons and low frequency alleles in our dataset, as it exhibited the lowest proportion of singletons among all SNP datasets. Yet, among all datasets, GATK had the highest medium site heterozygosity. It appears that GATK slightly overestimates mid-frequency alleles, because our genotype validations revealed that in 30% of the cases where GATK called a heterozygous and CLC a homozygous genotype, CLC was correct. Thus, our results suggest that with increasing minor allele frequency, GATK starts calling alternative alleles more aggressively due to the population prior in multi-sample analysis.

It appears that CLC generally underestimates heterozygosity. The CLC dataset consisted of an excess of singletons, suggesting that CLC called sequence errors as a genetic variants to a greater extent. Thus, the CLC dataset contained the lowest overall site heterozygosities among all datasets. Detailed investigation of discordantly called genotypes revealed that almost all of these genotypes were homozygous with CLC, but heterozygous with the other callers. To our surprise, CLC largely miscalled genotypes as heterozygous which were correctly assigned as homozygous by GATK (89% correct by GATK).

Our results indicate that SAMtools is more restrictive in SNP calling than GATK and CLC. The SAMtools dataset consisted of considerably fewer SNPs than GATK and CLC, but the degree of overlap with the other datasets was much higher than for the other datasets. SAMtools showed the highest tendency to assign heterozygous genotypes in cases of discordantly called genotypes among callers. For example, the few discordantly called genotypes were strongly biased in that 84% were heterozygous with SAMtools, but homozygous with GATK.

3.4 Discussion

Our study provides a framework for the generation of genome-wide SNP datasets for population genomic studies, from laboratory procedures to bioinformatics, which is widely applicable in non-model species. We present an improved protocol for highly efficient and more precise reduced genome complexity sequencing that simultaneously allows discovery of novel SNPs and genotyping. Using data generated from 31 wild-born orangutans from two populations, we observed significant inconsistencies among three commonly used SNP callers (*CLC Genomics Workbench, GATK Unified Genotyper and SAMtools*). These inconsistencies among the SNP datasets led to strong disagreement in outliers detected in scans for signatures of natural selection. This shows the potential impact on downstream biological analyses and emphasizes the need to critically evaluate the accuracy of SNP and genotype calling in population genomic studies.

We present a refined iRRL method presenting an improvement of the approach by van Tassel *et al.* (van Tassell *et al.* 2008). Several key modifications greatly enhanced the effectiveness of genotyping-by-sequencing, as measured by target sequence efficiency. Target sequence

efficiency was high because we focused on laboratory procedures to obtain homologous sequences across individuals, i.e. reproducible fragment generation and precise size selection. To our knowledge, these procedures do not seem to have received sufficient attention in the literature, probably because most studies pooled individuals to develop SNP markers (e.g. Wiedmann *et al.* 2008; van Bers *et al.* 2010; Kraus *et al.* 2011) without the direct aim of estimating allele-frequencies.

The importance of uniform fragment selection is well illustrated by our *in-silico* digests of the orangutan reference genome. An imprecise isolation of fragments would have led to a substantial change in the overall composition of fragment libraries across samples. This in turn would have caused a substantial increase in missing genotypes because of significantly reduced overlap of homologous fragments. Thus, accurate size selection and generation of uniform fragments to achieve high sequences homology are paramount in producing high-quality RRLs that maximize the amount of biological information.

The higher and more constant target sequence efficiencies for Sumatran West Alas individuals (median 97%) compared to Bornean South Kinabatangan individuals (median 86%) were most likely caused by carrying out the initial *in-silico* digest, which predicted our target sites, on the Sumatran reference genome. Since Sumatran and Bornean orangutans diverged more than 400,000 years ago, (Locke *et al.* 2011; Nater *et al.* 2011; Nater 2012), Bornean orangutans will inevitably exhibit more mutations at restriction sites.

We also improved previous RRL approaches by minimizing the loss of DNA during purification steps, thus facilitating single-sample library construction. Economical handling of DNA is particularly relevant when studying species for which sample quantity is a limiting factor, which is the case for most wild animal populations. A high DNA recovery rate during purification steps is especially important when dealing with low template amounts (<100 ng), where DNA loss will be disproportionately higher for technical reasons, and/or targeting only a small fraction of the genome. So far, these problems have been circumvented by pooling samples. Our DNA recovery rate of >95 % in the purification steps is considerably higher than obtained through conventional methods using extractions from gels and/or silica columns [<80%; QIAquick Spin Handbook Qiagen].

From a bioinformatical perspective, we demonstrate that different SNP callers lead to substantially different SNP datasets, in spite of applying rather conservative quality filters. For example, we applied a phred-scaled mapping and base quality threshold of ≥Q30, corresponding to an error probability of ≤0.1%. In contrast, other studies only apply Q20 (1.0% error probability) (e.g. van Bers *et al.* 2010; Amaral *et al.* 2011; Nielsen *et al.* 2011; Jonker *et al.* 2012). Furthermore, our median sequence coverage of 41x (minimal cut-off of 10 reads) is substantially higher than that found in other studies, in which sequencing depth is usually between 6-16x with lower cut-off values than used in this study (e.g. Wiedmann *et al.* 2008; Kerstens *et al.* 2009; Sanchez *et al.* 2009; Esteve-Codina *et al.* 2011; Kraus *et al.* 2011).

There are three main reasons for the conspicuous differences among the SNP datasets. First, the SNPs dropping out because of our population-based filtering were different among the GATK, SAMtools and CLC datasets. Second, although we used identical mapped short reads and filtering criteria on the raw data to call SNPs and genotypes, we cannot exclude a potential influence of the poor F5 sequence read performance due to specific internal filters of SNP callers. Third and most importantly, some differences will arguably be related to the conceptually very different methods of SNP identification and genotype assignment (Li *et al.* 2009; McKenna *et al.* 2010; DePristo *et al.* 2011; Nielsen *et al.* 2011).

Intersect strategies have been proposed to reduce caller-specific errors (e.g. Nielsen *et al.* 2011). The estimated genotype accuracy of 96.5% of intersected genotypes is higher than in comparable studies that use only one caller (e.g. 47-84% (Sanchez *et al.* 2009; van Bers *et al.* 2010; Everett *et al.* 2011); 89-95% (van Tassell *et al.* 2008; Hyten *et al.* 2010; Kraus *et al.* 2011; Sharma *et al.* 2012b)). Yet, most of these studies actually only verified the polymorphic state of SNPs but not individual genotype calls. Thus, the true genotype error rate in these studies is almost certainly higher than estimated.

The intersect strategy seems to be appealing because false-positive assignments should be minimized. However, it is inevitably less sensitive towards SNP discovery (Nielsen *et al.* 2011). The appropriate strategy and filter stringencies for each study depend on the specific needs of downstream analyses. Nonetheless, apart from higher false-negatives rates, as observed in our dataset, intersecting genotype calls might also introduce non-random biases. More detailed investigations will be required to fully appreciate the consequences of intersecting strategies.

Among all datasets, the general patterns tend to agree with previous detailed studies on orangutan population genetics and demographic history. For instance, the higher number of singletons and low-frequency alleles we observe in the Sumatran West Alas population is in agreement with previous studies using conventional genetic markers (mitochondrial DNA, microsatellites) (Steiper 2006; Nater *et al.* 2011; Nater *et al.* 2013). Furthermore, the slightly higher site heterozygosities in the South Kinabatangan population are also in agreement with previous studies using conventional genetic markers (Kanthaswamy & Smith 2002; Goossens *et al.* 2005; Sharma *et al.* 2012a; Nater *et al.* 2013).

Many downstream analyses in population genomics, such as selection tests or demographic inferences rely on the allele-frequency spectra (Nielsen 2005). Thus, biological conclusions drawn from such analyses may well change depending on which SNP caller has been used. This possibility is illustrated by the extremely low overlap of identified outlier regions in our sliding-window analyses to detect selective sweeps based on population differentiation.

Apart from reliable SNP analysis, the accurate characterization of the allele-frequency spectra is mainly influenced by three sources of bias. First, allele frequencies will not be representative of the population if there is a sampling bias (Excoffier *et al.* 2009; Garvin *et al.*

2010; Schoville *et al.* 2012). To address this issue and reduce this bias, we carefully selected study animals and verified population origins. By contrast, genomic studies often rely on zoo animals with unknown population provenance (if wild-born) or apply a limited sampling schema (e.g. Locke *et al.* 2011), and thus there are likely inherent sampling biases.

Second, the discovery of SNPs in a subset of individuals for subsequent genotype calling in an extended sample set will lead to ascertainment bias (Helyar *et al.* 2011; Seeb *et al.* 2011a; Seeb *et al.* 2011b). The degree of ascertainment bias depends on the representativeness of the sampling scheme of individuals used for the initial SNP discovery (Garvin *et al.* 2010). Especially in population and conservation genomics, ascertainment bias is a serious problem when assessing, for instance, genetic diversity. Low-frequency variants will be underestimated and a systematic bias will be introduced (Helyar *et al.* 2011). The key strength of reduced genome complexity approaches is that this form of ascertainment bias can be minimized by the genotyping-by-sequencing principle.

Third, it is biologically relevant to also capture rare alleles, which is the reason why we established individual libraries (i.e. no pooling of samples). Low-frequency alleles are important in estimates of demographic parameters (e.g. Thornton & Andolfatto 2006) and studies of positive (Przeworski 2002) and purifying selection (Charlesworth *et al.* 1993).

The framework provided in this study will be valuable to generate genome-wide SNP datasets in the emerging fields of population, conservation and landscape genomics. Our iRRL protocol is part of a growing suite of sequencing methods, which have completely changed study designs and hold great promise for studies of ecology and evolution in diverse species. The strength of reduced-genome-complexity RRL methods is that they can be applied to any DNA-based life form, opening up the field of population genomics to smaller research groups studying organisms for which large-scale genetic data is not yet available. Until high-throughput sequencing becomes more affordable and bioinformatical advances allow routine whole-genome re-sequencing of populations, we expect that reduced-genome-complexity approaches will remain essential for population genomic studies particularly in non-model organisms with large genomes.

Conclusions

We generated SNP datasets for 31 wild-born orangutans from two populations representing the first effort of large-scale SNP discovery and genotyping of orangutans with known population provenance. In the field of population genomics, researchers need to exert caution when generating genome-wide SNP datasets. We show that accurate generation of homologues fragments in reduced-genome-complexity sequencing is paramount, especially for pooled samples with no control for missing genotypes in the estimation of allele frequencies. We present an improved RRL protocol (iRRLs), which allows sampling only a fraction of the genome with maximized sequence overlap among individuals. The scale and efficiency achieved with our iRRL protocol demonstrates its suitability to generate genome-

wide SNP datasets. Our direct comparison of three popular SNP callers demonstrated that depending on the calling algorithm, sequence depths and filtering criteria, substantially different SNP datasets are obtained that will affect downstream analyses and thus might have a substantial effect on biological conclusions. When only applying a single SNP caller, we advise to use a probabilistic algorithm and call genotypes in a multi-sample mode. In our study, the Bayesian framework of the *Unified Genotyper* of the *GATK* showed a higher sensitivity in discovering SNPs than the framework of *SAMtools* with similar genotype calling accuracy.

3.5 Methods

DNA samples

We sampled two orangutan populations, one from Borneo and one from Sumatra (Figure 1). To obtain sufficient amounts of high-quality DNA, we collected blood samples from rehabilitant wild-born orangutans. We sampled 15 individuals from the West Alas population (WA, *Pongo abelii*, northwestern Sumatra) at the Batu Mbelin Quarantine Center of the Sumatran Orangutan Conservation Programme, and 16 individuals from the South Kinabatangan population (SK, *Pongo pygmaeus morio*, northeastern Borneo) at the Sepilok Orangutan Rehabilitation Centre, Shangri-La's Rasa Ria Resort Sanctuary and Lok Kawi Wildlife Park in Sabah. Whole blood samples were taken during routine veterinary examinations and stored in EDTA blood collection tubes at -20°C. The collection and transport of samples were conducted in strict accordance with Malaysian, Indonesian and international regulations. Samples were exported from Malaysia and Indonesia to Switzerland under the Convention on International Trade of Endangered Species in Fauna and Flora (CITES) permit numbers 4872/2010 (Sabah, Malaysia) and 06968/IV/SATS-LN/2005 (Indonesia), respectively. Detailed information on the sampled individuals is provided in Table S1. We verified the individual's population origin by genetic assignment tests and Bayesian clustering algorithms as described in the Supplementary Information.

To minimize DNA shearing, we avoided repeated thawing and freezing of samples and used only wide-bore tips and avoided vortexing during DNA extraction. Genomic DNA was extracted using the Gentra Puregene Kit (Qiagen) according to the manufacturer's instructions, including RNase treatment, but with the following modifications for clotted blood: we added twice the amount of Cell Lysis Solution as well as 7 µl of Proteinase K (20mg/ml, Promega) per 100 mg blood clot to the samples, followed by incubation for 3 hours at 55°C in a slowly revolving overhead rotator. If the solution still appeared to be viscous after this treatment, we increased incubation time and added more Proteinase K as required until complete liquefaction. We also used twice the recommended amount of Protein Precipitation Solution and incubated samples on ice for 10 minutes after addition of the solution to promote protein precipitation. DNA pellets were eluted in ddH_2O instead of

DNA Hydration Solution (Qiagen) to facilitate DNA concentration using a SpeedVac vacuum centrifuge (Savant).

Reduced-representation libraries construction

We performed *in-silico* digests of the orangutan reference genome (*ponAbe2* (Locke *et al.* 2011)) to evaluate a suitable restriction enzyme to construct iRRLs using custom perl scripts. We tested 23 commercially available Type II DNA blunt-end cutters (4-6 bp recognition sites) in multiple combinations (Table S7). Selection criteria were: (i) target size range 70-200 bp, (ii) number of fragments predicted in size range corresponding to ~1% of the genome, and (iii) low repetitive element content. We chose *Hae*III because in the size range of 104-123 bp, *Hae*III did not produce obvious repetitive elements based on visual inspection of the fragment distribution profile (Figure 2), and covered ~1% of the genome. The enzyme *Hae*III has also been selected in previous studies (van Tassell *et al.* 2008; Hyten *et al.* 2010; Jonker *et al.* 2012), and thus might be a good candidate enzyme for reduced-genome-complexity sequencing in general.

In cases where there is no reference genome available, the evaluation for a suitable enzyme could also be carried out in the laboratory, for example by analyzing the fragment distribution of digested genomic DNA using high resolution electrophoresis (e.g. Agilent 2100 Bioanalyzer). These instruments offer tools to estimate the represented genome proportion of fragments within a given size range.

We established iRRLs for each individual by digesting 20 µg of genomic DNA with 200 units of *Hae*III (50,000 U/ml, New England Biolabs) in a total volume of 32 µl. Digests were run on high-resolution Spreadex EL400 Wide Mini S-2x13 gels with M3 size marker in a SEA 2000 electrophoresis chamber (all Elchrom Scientific, Switzerland) in 1x TAE buffer at 120 Volt for 147 min, keeping temperature constant at 55°C to ensure reproducibility of fragment migration. The running time was the evaluated optimum for the target size range using the ELQuant Software (www.elchrom.com). Each digest was equally distributed in two separate wells to avoid DNA overloading. We stained gels with GelRed (Biotium) and excised fragments between 104 bp and 123 bp on a UV-transilluminator using a long-bladed sharp knife, keeping UV exposure as short as possible.

DNA fragments were recovered by electro elution to achieve high DNA recovery rate (>95%). For this, we prepared dialysis membranes (Carl Roth, 1785.1 Dialysierschlauch Visking) of approximately 5 cm width, which we sealed on one side with a plastic clip (Carl Roth, H277.1 Verschlussklammer). We filled each dialysis membrane with 1 ml of 1x TAE buffer and placed gel slices in the membrane in the same running orientation as in the electrophoresis run (illustrated in Figure S3). We closed the dialysis membrane with a second plastic clip and avoided trapping any air bubbles inside the membrane. Packages were then placed in an SEA 2000 electrophoresis chamber filled with 1x TAE buffer. We applied 90 Volts for 100 minutes, followed by 1 minute of reverse polarity to detach DNA from the wall of the membrane. We

gently massaged the packages to mix the eluted DNA in the buffer. After this, we carefully opened one of the clips to gently pipet out the buffer containing the eluted DNA. The DNA was purified using the MinElute PCR Purification Kit (Qiagen). This way, we obtained between 2 and 20 ng of DNA per sample. Individual barcoding of iRRLs and SOLiD sequencing library preparation was performed according to the SOLiD ChiP-Seq protocol step 11 (Applied Biosystems, 2010), which had been optimized for low template quantities (e.g. Agencourt AMPure XP beads for purification steps). We restricted library amplification to six PCR cycles only, so as to minimize the risk of over-amplification. After library quality control on an Agilent Bioanalyzer 2000, we normalized samples and sent pooled libraries to the Functional Genomics Center Zurich, Switzerland (FGCZ) for sequencing on a SOLiD 4^{TM} System with paired-end (50/35) chemistry (Life Technologies).

SNP discovery and genotype calling

Raw sequence reads were processed and mapped to the orangutan reference genome *ponAbe2* (Locke *et al.* 2011) using the SOLiD LifeScope v.2.5.1 package (Life Technologies) according to their guidelines. We used Picard v.1.57 [http://picard.sourceforge] to merge mapping files for each individual from different SOLiD runs and adjust read group headers. We called SNPs using three different programs as described below.

We performed simultaneous multi-sample SNP and genotype calling with the Unified Genotyper of the GATK v.2.5-0 (McKenna *et al.* 2010; DePristo *et al.* 2011) with the following thresholds: phred-scaled mapping and base qualities ≥ 30 ('-mmq 30 -mbq 30'). We filtered out low-quality genotypes (GQ<10) and genotypes covered by less than 10 or more than 1000 reads ('-minGQ 10 -minDP 10 -maxDP 1000') using VCFtools v.0.1.9 [70]. Sites which were homozygous after this filtering were removed. Finally, we disregarded sites with more than two alleles and only retained sites with a genotype call for a minimum of eight individuals per population that had passed all quality filters applying custom R scripts.

As a second probabilistic caller, we used SAMtools v.0.1.19 (Li *et al.* 2009) to call SNPs and genotypes in all individuals simultaneously. We applied the same filter thresholds as for the GATK dataset and used defaults settings otherwise (except for deactivating the base alignment quality realignment with the -B parameter: 'samtools mpileup -q 30 -Q 30 -B'). Post-filtering of SNP and genotype calls was conducted as for the GATK dataset.

As an alternative non-probabilistic approach, we discovered SNPs with the quality-based variant detection tool of the CLC Genomics Workbench v.5.0.1 (CLC bio) following the same quality requirements as applied in the GATK/SAMtools calls. Since the CLC version we used did not offer multi-sample calling (i.e. analyzing all individuals simultaneously) at the time of this study, we detected SNPs for each individual separately and merged the SNP data subsequently using R scripts. In this merged dataset, a missing call for an individual for a certain SNP position could arise either because this individual is homozygous for the reference allele or because this site was not sequenced. To obtain this information for all

missing genotypes, we used SAMtools v.0.1.12a (Li *et al.* 2009). We called genotypes according to common practice, applying fixed cut-off rules based on read counts (Nielsen *et al.* 2011) with *ad-hoc* R scripts. Sites with an alternative allele frequency between 0-15% were called homozygous for the reference allele, sites with an alternative allele frequency between 20-80% as heterozygous, and sites between 85-100% alternative allele frequency as homozygous for the alternative allele. To be conservative, we denoted sites with borderline alternative allele frequencies (i.e. 15-20% and 80-85%) as 'N'. We only accepted sites with a maximum of two alleles and covered by minimal eight individuals per study population, as we had done for the GATK and SAMtools datasets.

Finally, we used custom R scripts to intersect the GATK, SAMtools and CLC genotype calls for each individual at all sites (without the population-based filters) only retaining identical genotype calls. After merging the individual data, we again excluded sites with more than two alleles and genotypes in less than eight individuals per population as performed with the other datasets (population-based filtering). Note that further filters could be applied for SNP and genotype calling from high-throughput sequencing data such as filtering clusters of SNPs (for a list see Supporting Information of Auton *et. al.* (Auton *et al.* 2012))

SNP and genotype validation

To assess genotype accuracy between the probabilistic callers and CLC and estimate the error rate of identical genotype calls, we validated 180 genotypes by classical Sanger sequencing. We randomly picked 58 overlapping SNPs with the only requirements that a minimum of 10 individuals per population had a genotype called, and that at least one individual showed a conflicting genotype call. We validated genotypes of several individuals at those SNPs, which appeared to belong to three different classes: (i) homozygote genotype with GATK/SAMtools but heterozygote with CLC, (ii) heterozygote genotype with GATK/SAMtools but homozygote with CLC, (iii) identical genotype call. We also validated (iv) singleton sites (only one alternative allele called in the entire dataset) that were determined by only one of the callers through Sanger-sequencing of the individual that exhibited the singleton. Additionally, to specifically investigate the genotype accuracy of SNPs present in the intersect data of all three callers, we randomly picked an additional 14 SNPs from this dataset. BEDtools v.2.16.2 (Quinlan & Hall 2010) was used to extract the DNA sequences 400 bp downstream and upstream of the targeted SNPs from the orangutan reference genome *ponAbe2*. PCR primers flanking the SNPs were designed with Primer3 (Rozen & Skaletsky 1999) (Table S6). We verified genotypes by sequencing PCR products on a 3730 DNA Analyzer (Applied Biosystems). Details on PCR conditions, cycle sequencing and data analyses are provided in the Supporting Information.

Statistical analyses

We considered all sites with mapping and base quality phred scores of ≥ 30, and a minimal sequence depth of 10 as high-quality base pairs (bp_{hiqual}). We estimated the target efficiency

of our iRRL protocol by calculating which percentage of the actually sequenced bp_{hiqual} was predicted by our *in-silico* digest of *ponAbe2* with *Hae*III (=target sites). Furthermore, for each SNP and population we calculated the observed site heterozygosity as the number of individuals carrying both alleles divided by the total number of called genotypes in this population. Kernel density plots of the minor allele frequency and site heterozygosity distributions were drawn in R with the 'sm' package (Bowman & Azzalini 2010). We assessed the significance of equality of the density estimates among the different datasets with the 'sm.density.compare' function with 10,000 permutations.

In addition, we performed sliding-window analyses for each dataset to detect selective sweeps in the genome based on population differentiation using custom R-scripts. For all SNPs we estimated population differentiation using allele-frequency differentials, defined as: $D = \sum[abs(\mathrm{pSK} - \mathrm{pWA}) + \mathrm{abs}(\mathrm{qSK} - \mathrm{qWA})]/2$, where p and q denote the frequencies of the two alleles for each SNP. We scanned each chromosome ('chrXY_random' excluded) and calculated for each window (100 kb window size, 25 kb step size) the average *D* of all SNPs. We arbitrarily defined outlier regions as windows with an average population differentiation $D > 0.95$ (covered by at least 2 SNPs).

Acknowledgments

We thank the Functional Genomics Centre Zurich and Gerrit Kuhn from Life Technologies for their support on sequencing and bioinformatics. We are grateful to Giada Ferrari for assisting in the CLC SNP analyses, David Marques who genotyped some of the WA individuals, Glauco Camenisch for help with *in-silico* digests, and Erik Willems for providing the map of Borneo and Sumatra. We thank Christian Lexer for general support. We are indebted to the staff at the Sepilok Orangutan Rehabilitation Centre, the Shangri-La's Rasa Ria Resort, the Lok Kawi Wildlife Park and the Sumatran Orangutan Conservation Programme who helped collecting samples. Furthermore, we thank the following institutions for supporting our research: Sabah Wildlife Department (SWD), Indonesian State Ministry for Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), and Leuser International Foundation (LIF). Financial support was provided by the Forschungskredit University of Zurich (to MPG), A.H. Schultz Foundation (to MK and MPG), Swiss National Science Foundation (grant no. 3100A-116848 to MK and CPvS), Julius-Klaus Foundation (to MK), Leakey Foundation (to MPG), and the Anthropological Institute & Museum at the University of Zurich.

Author Contributions

MPG, MK conceived and coordinated the study with input from CPvS. MPG designed and performed experiments. KNS, AN, and RHSK contributed to the experimental design. BG, MPG, MK, RS, IS, LNA, LC, and CPvS provided genetic samples. AN and NA supported the population genetic assessment of study individuals. BN contributed ideas and reagents. RB

and MPG conducted in-silico analyzes. AP contributed reagents and performed sequencing. RB carried out short read mapping. MPG performed bioinformatical analyses. KNS supported the bioinformatical analyses. MPG and MK wrote the manuscript. NA edited the manuscript. KNS, AN, BN, RHSK, LC, RB, RS, BG, and CPvS commented on the manuscript. All authors read and approved the final manuscript.

3.6 Supporting Information

The raw sequence data is available in the NCBI Sequence Read Archive, BioProject: PRJNA230877; BioSamples: SAMN02439270-SAMN02439300.

Population genetic assessment of study individuals

Our detailed knowledge of the orangutan population structure through previous studies (Nater *et al.* 2011; Arora *et al.* 2012; Nater 2012; Nietlisbach *et al.* 2012; Nater *et al.* 2013) on both Borneo and Sumatra, allowed us to assign individuals to their population of origin. We genotyped the animals of the current study at the same 27 highly polymorphic microsatellite makers used in previous studies (Arora *et al.* 2010; Nietlisbach *et al.* 2010). We then used the program STRUCTURE v2.3.3 (Pritchard *et al.* 2000) to analyze the study animals for this study together with genotype data from 219 individuals from previous studies (Nater *et al.* 2011; Arora *et al.* 2012; Nater 2012; Nietlisbach *et al.* 2012; Nater *et al.* 2013) that represent all major genetic clusters on both islands. For each individual we estimated the membership coefficient Q (Pritchard *et al.* 2000) of belonging to a particular cluster (Figure S1). Details on the analysis and the general observed population structure can be found in Nater (Nater 2012). The highest hierarchical level clearly separates Bornean and Sumatran individuals. Analyzing each island separately, three distinct clusters seem to best describe the structure observed in the variation analyzed on Sumatra and five clusters on Borneo. All individuals from this study show high membership to the South Kinabatangan and West Alas cluster, which had also been confirmed by phylogenetic analyses of mitochondrial DNA gene sequences in Nater et. al. (Nater *et al.* 2011) .

SNP validation

All PCR reactions were performed in a 10 µl reaction volume containing 5 ng of template DNA, 0.1 µM of each primer, 0.2 mM dNTPs, 1 x Phire PCR Buffer and 0.1 units Phire Hot Start II DNA Polymerase (both Finnzymes). PCR conditions were as follows: initial denaturation at 98°C for 30 s, followed by 35 cycles of 98°C for 5 s, 62°C for 10 s, 72°C for 15 s, and final extension at 72°C for 2 min. Direct cycle sequencing was performed with 0.5 µl PCR product in a 10 µl reaction volume containing 1x sequencing buffer (400 mM Tris, 10 mM $MgCl_2$, pH 9.0), 0.4 µM forward primer and 0.3 µl BigDye Terminator v3.1 (Life Technologies). Cycle sequencing conditions were as follows: initial denaturation at 95°C for 45 s, 30 cycles of 95°C for 30 s, 52°C for 20 s, 60°C for 2 min. Samples were sequenced on a 3730 DNA Analyzer (Life Technologies). We aligned generated sequences with the reference genome sequence *ponAbe2* (Locke *et al.* 2011) using the SeqMan program of the Lasergene 8 software package (DNASTAR) and visually called genotypes for the target SNP position based on the trace data.

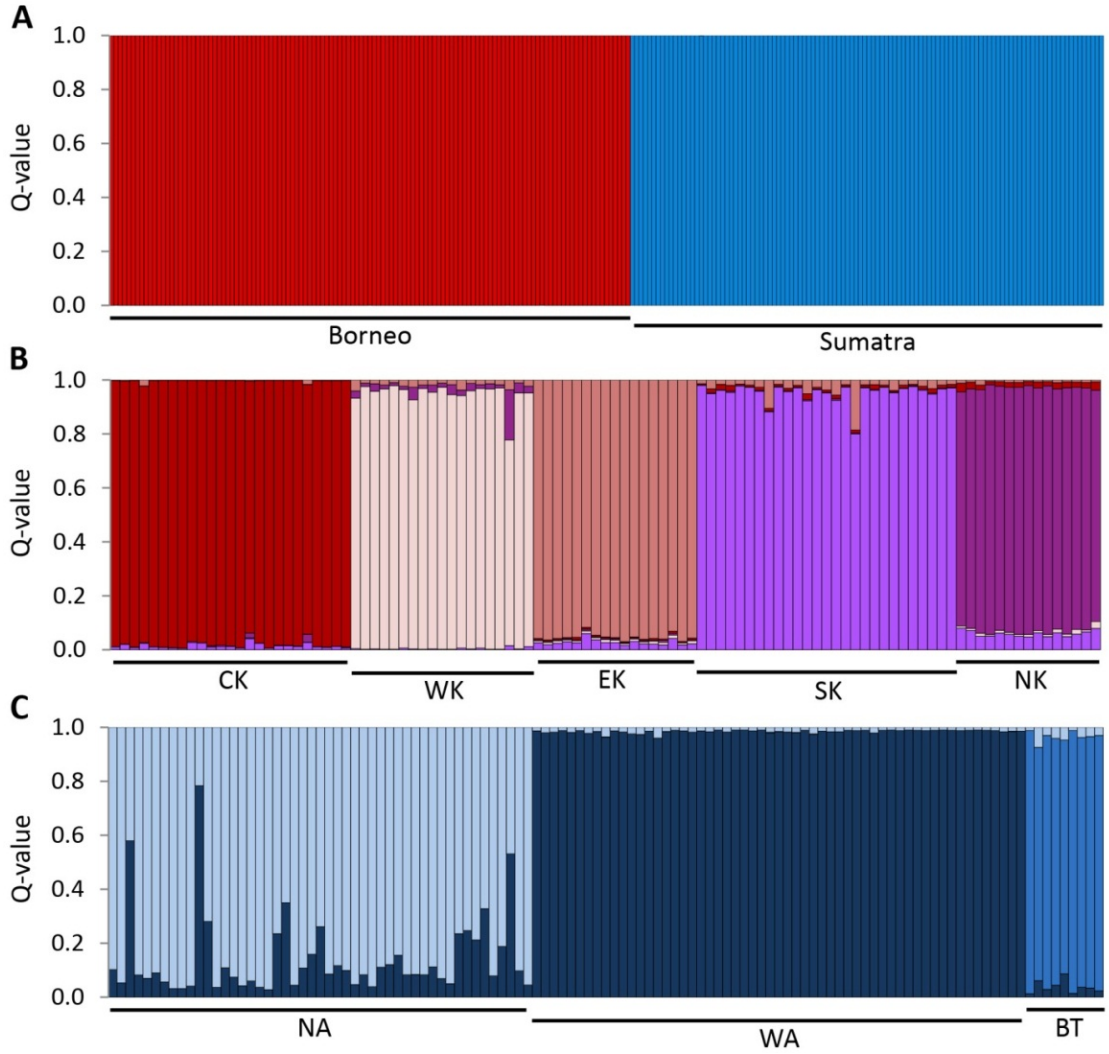


Figure S1. STRUCTURE analyses to identify the population origin of study individuals. The analyses were based on 27 microsatellite markers. The membership coefficients Q are average values over ten iterations with the same model parameters (admixture model with burn-in of 3x105 followed by 3x106 MCMC steps). Each bar represents a single individual. (A) Bornean and Sumatran individuals separate at the highest hierarchical level. (B) Within Borneo (n=104) we observe five clusters (most likely number of clusters as inferred by Arora *et al.* (Arora *et al.* 2010) and Nater (Nater 2012)). Study individuals (n=16) were sampled from the South Kinabatangan cluster. CK: Central Kalimantan (n=25), WK: West Kalimantan (n=20), EK: East Kalimantan (n=17), SK: South Kinabatangan (n=25), NK: North Kinabatangan (n=17). (C) Within Sumatra (n=115) we observe three clusters (most likely number of clusters as inferred by Nater *et al.* (Nater *et al.* 2013) and Nater *et al.* (Nater 2012)). Study individuals (n=15) were sampled from the West Alas population. NA: North Aceh and Langkat (n=49), WA: West Alas (n=57), BT: Batang Toru (n=9).

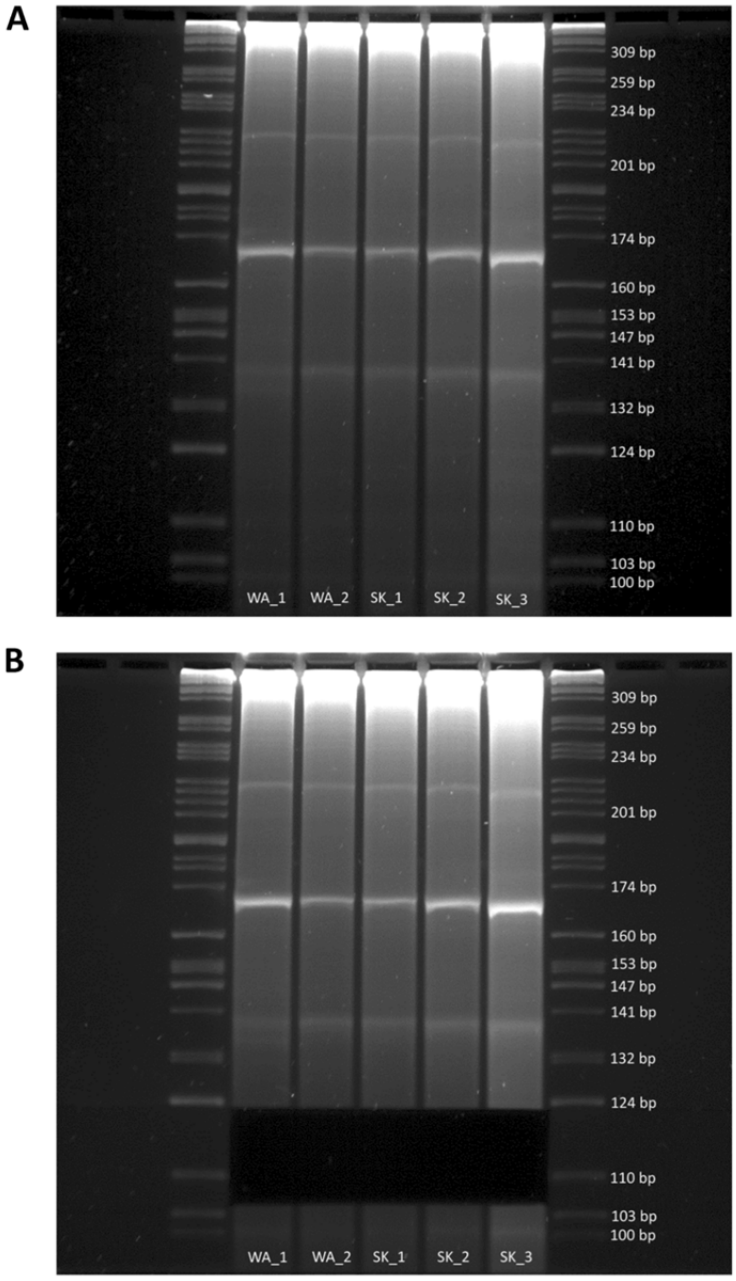


Figure S2. Example image of high precision excision of target fragments from Spreadex gel. HaeIII digested genomic DNA was separated on high-resolution Spreadex EL400 Wide Mini S-2x13 gels with M3 size marker. Bands at e.g. 101 bp, 136 bp, and 170 bp represent repetitive elements as predicted by the in-silico HaeIII digest of the orangutan reference genome (see Figure 2). (A) Image of a gel prior to fragment excision. (B) Image of the same gel after excision of DNA fragments in the target size range 104-123 bp.

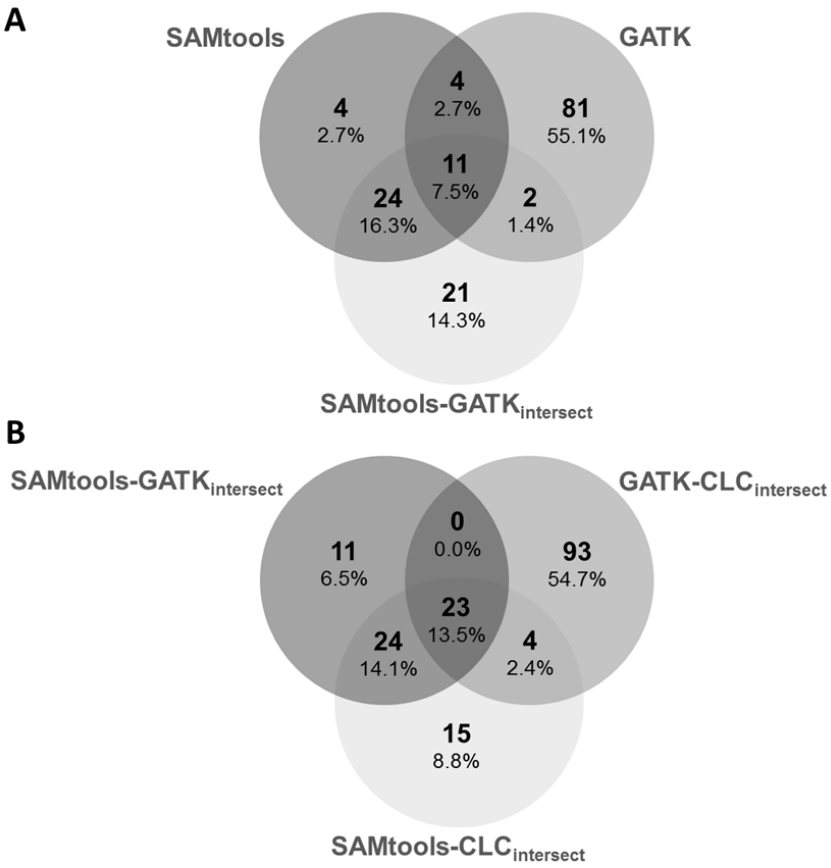


Figure S3. Overlap of outlier regions among SNP datasets in genome-wide scans for positive selection. For all SNP datasets we performed sliding-window analyses (100 kb window, 25 kb step size) of the absolute allele-frequency differential (D) between the South Kinabatangan and West Alas population. All windows with an average window D > 0.95 were considered as outliers, i.e. candidate regions for selective sweeps. (A) Comparison of SAMtools and GATK with their intersect dataset. (B) Comparison of the three intersect datasets.

Table S1. List of study individuals.

Individual	Population	Species	Region	Sex	Sample Type	Sampling Date
WA_1	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	21.11.2005
WA_2	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	21.11.2006
WA_3	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	21.11.2007
WA_4	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	24.12.2005
WA_5	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	12.11.2005
WA_6	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	24.12.2005
WA_7	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	08.10.2005
WA_8	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	08.10.2006
WA_9	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	06.05.2006
WA_10	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	24.12.2005
WA_11	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	06.09.2006
WA_12	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	04.02.2006
WA_13	West-Alas	*Pongo abelii*	Northwest Sumatra	M	Whole Blood	06.05.2006
WA_14	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	12.11.2005
WA_15	West-Alas	*Pongo abelii*	Northwest Sumatra	F	Whole Blood	04.02.2006
SK_1	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	F	Whole Blood	04.02.2010
SK_2	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	F	Whole Blood	04.02.2010
SK_3	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	05.02.2010
SK_4	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	05.02.2010
SK_5	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	06.02.2010
SK_6	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	06.02.2010
SK_7	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	06.02.2010
SK_8	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	06.02.2010
SK_9	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	06.02.2010
SK_10	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	F	Whole Blood	06.02.2010
SK_11	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	08.02.2010
SK_12	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	08.02.2010
SK_13	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	08.02.2010
SK_14	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	M	Whole Blood	16.02.2010
SK_15	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	F	Whole Blood	16.02.2010
SK_16	South Kinabatangan	*Pongo pygmaeus morio*	Northeast Borneo	F	Whole Blood	16.02.2010

Table S2. Basic sequencing and mapping statistics for all study individuals.

Individual	Population	Total beads	Mapped reads F3	Mapped reads F5	Total mapped reads	% mapped F3	% mapped F5	Total bp passing high quality filters (bp_{hiqual})[a]	bp_{hiqual} matching *in-silico* predictions[b]	iRRL target efficiency (%)[c]	Mean sequence coverage[d]	Median sequence coverage[d]
WA_1	WA	49,292,298	37,470,056	3,615,166	41,085,222	76.0	7.3	12,673,892	12,357,484	98	70	41
WA_2	WA	37,174,228	28,664,287	2,370,716	31,035,003	77.1	6.4	8,948,586	8,657,259	97	69	42
WA_3	WA	46,917,923	35,953,417	3,773,266	39,726,683	76.6	8.0	10,930,563	10,545,929	96	72	42
WA_4	WA	33,299,284	24,009,171	1,914,332	25,923,503	72.1	5.7	6,346,072	6,179,537	97	69	43
WA_5	WA	39,263,916	29,391,148	2,954,029	32,345,177	74.9	7.5	10,845,558	10,560,344	97	72	42
WA_6	WA	46,105,067	34,168,912	3,607,637	37,776,549	74.1	7.8	11,628,279	11,185,593	96	71	41
WA_7	WA	32,229,848	24,045,574	2,001,818	26,047,392	74.6	6.2	11,628,279	11,185,593	96	71	41
WA_8	WA	31,665,241	23,601,524	2,562,180	26,163,704	74.5	8.1	8,769,047	8,570,985	98	75	43
WA_9	WA	92,913,519	42,386,737	3,667,274	46,054,011	45.6	3.9	12,854,122	11,192,488	87	56	41
WA_10	WA	34,009,597	25,857,283	2,274,126	28,131,409	76.0	6.7	8,169,085	7,959,777	97	71	43
WA_11	WA	20,362,438	15,793,068	1,361,670	17,154,738	77.6	6.7	8,008,214	7,804,744	97	69	43
WA_12	WA	73,041,931	56,023,657	5,360,192	61,383,849	76.7	7.3	13,747,885	13,271,073	97	70	41
WA_13	WA	58,385,777	43,242,808	3,897,908	47,140,716	74.1	6.7	13,146,186	12,761,927	97	70	41
WA_14	WA	49,821,551	38,891,523	3,909,350	42,800,873	78.1	7.8	12,491,163	12,127,905	97	71	41
WA_15	WA	30,813,183	22,984,433	2,328,673	25,313,106	74.6	7.6	8,187,784	8,005,757	98	74	43
Total	**WA**	**675,295,801**	**482,483,598**	**45,598,337**	**528,081,935**	**/**	**/**	**/**	**/**	**/**	**/**	**/**
Mean	**WA**	**45,019,720**	**32,165,573**	**3,039,889**	**35,205,462**	**73.5**	**6.9**	**12,407,006**	**10,157,760**	**96**	**70**	**42**
Median	**WA**	**39,263,916**	**29,391,148**	**2,954,029**	**32,345,177**	**74.9**	**7.3**	**10,930,563**	**10,560,344**	**97**	**71**	**42**
SK_1	SK	53,203,171	33,946,147	8,667,237	42,613,384	63.8	16.3	19,172,077	17,342,891	90	50	41
SK_2	SK	44,278,780	16,014,503	4,075,314	20,089,817	36.2	9.2	13,637,944	6,422,988	47	47	40
SK_3	SK	51,299,839	35,269,730	9,020,858	44,290,588	68.8	17.6	17,935,765	16,158,644	90	51	41
SK_4	SK	34,618,999	28,060,571	7,533,288	35,593,859	81.1	21.8	14,668,357	13,496,898	92	53	41
SK_5	SK	55,926,738	40,069,230	10,302,123	50,371,353	71.6	18.4	18,070,791	16,863,737	93	51	41
SK_6	SK	47,321,982	28,061,965	7,041,823	35,103,788	59.3	14.9	19,106,920	16,085,693	84	50	41
SK_7	SK	47,155,939	35,209,706	9,395,908	44,605,614	74.7	19.9	17,643,456	14,945,703	85	52	41
SK_8	SK	48,113,617	29,306,604	7,639,421	36,946,025	60.9	15.9	19,644,644	16,560,683	84	50	41
SK_9	SK	49,472,477	40,449,154	10,599,426	51,048,580	81.8	21.4	18,523,921	17,287,993	93	50	41
SK_10	SK	52,095,725	40,859,054	10,708,334	51,567,388	78.4	20.6	19,055,494	17,795,413	93	50	41

Table S2 (Continued)

Individual	Population	Total beads	Mapped reads F3	Mapped reads F5	Total mapped reads	% mapped F3	% mapped F5	Total bp passing high quality filters (bp_{hiqual})[a]	bp_{hiqual} matching *in-silico* predictions[b]	iRRL target efficiency (%)[c]	Mean sequence coverage[d]	Median sequence coverage[d]
SK_11	SK	55,678,462	36,679,163	9,408,008	46,087,171	65.9	16.9	18,021,511	13,289,969	74	48	40
SK_12	SK	51,018,856	36,594,500	23,157,125	59,751,625	71.7	45.4	20,047,467	12,174,784	61	49	40
SK_13	SK	30,373,598	9,980,894	2,528,617	12,509,511	32.9	8.3	8,593,406	3,816,017	44	50	40
SK_14	SK	44,086,074	24,160,878	6,334,668	30,495,546	54.8	14.4	18,302,919	15,975,745	87	50	41
SK_15	SK	50,425,043	32,272,868	8,373,266	40,646,134	64.0	16.6	18,364,849	15,663,597	85	51	41
SK_16	SK	47,164,781	35,798,036	9,403,785	45,201,821	75.9	19.9	17,593,190	15,949,358	91	51	41
Total	**SK**	**762,234,081**	**502,733,003**	**144,189,201**	**646,922,204**	**/**	**/**	**/**	**/**	**/**	**/**	**/**
Mean	**SK**	**47,639,630**	**31,420,813**	**9,011,825**	**40,432,638**	**65.1**	**18.6**	**22,396,785**	**14,364,382**	**81**	**50**	**41**
Median	**SK**	**48,793,047**	**34,577,927**	**8,844,048**	**43,451,986**	**67**	**17**	**18,186,855**	**15,962,552**	**86**	**50**	**41**

[a]Total number of base pairs sequenced passing all high quality filters such as mapping and base qualities ≥ 30 and sequence coverage ≥ 10 (sites used for SNP detection)

[b]Total number of high-quality base pairs sequenced which were predicted by *in-silico* digest of the orangutan reference genome (*PonAbe2*, Sumatra) using *Hae*III (=target sites)

[c]Note that the values are lower for SK individuals because the reference genome used for the *in-silico* digest stems from a Sumatran individual

[d]coverage of the bp_{hiqual} sites estimated with the *GATK Unified Genotyper*

Table S3. Comparison of the SNP and genotype calling of GATK and CLC for each individual.

Individual	Population	Genotype call in both	% same genotype called	% different genotype called
WA_1	WA	61488	97.14	2.86
WA_2	WA	33974	96.84	3.16
WA_3	WA	55734	96.84	3.16
WA_4	WA	21382	96.67	3.33
WA_5	WA	52464	96.92	3.08
WA_6	WA	58382	96.92	3.08
WA_7	WA	35303	97.48	2.52
WA_8	WA	39744	97.08	2.92
WA_9	WA	56767	93.59	6.41
WA_10	WA	31812	96.94	3.06
WA_11	WA	26593	96.67	3.33
WA_12	WA	76242	96.62	3.38
WA_13	WA	67988	97.08	2.92
WA_14	WA	66124	96.84	3.16
WA_15	WA	33824	97.51	2.49
Mean	**WA**	**47855**	**96.74**	**3.26**
Median	**WA**	**52464**	**96.92**	**3.08**
SK_1	SK	119674	98.34	1.66
SK_2	SK	81939	98.06	1.94
SK_3	SK	115359	98.38	1.62
SK_4	SK	92499	98.07	1.93
SK_5	SK	113192	98.29	1.71
SK_6	SK	120304	98.29	1.71
SK_7	SK	108940	98.34	1.66
SK_8	SK	120448	98.21	1.79
SK_9	SK	112495	98.24	1.76
SK_10	SK	116976	98.28	1.72
SK_11	SK	110325	97.09	2.91
SK_12	SK	123002	98.34	1.66
SK_13	SK	59343	97.88	2.12
SK_14	SK	120543	97.98	2.02
SK_15	SK	114463	98.38	1.62
SK_16	SK	110068	98.25	1.75
Mean	**SK**	**108723**	**98.15**	**1.85**
Median	**SK**	**113828**	**98.27**	**1.73**
Mean Overall		**79271**	**97.47**	**2.53**
Median Overall		**76242**	**97.51**	**2.49**

Table S4. Comparison of the SNP and genotype calling of GATK and SAMtools for each individual.

Individual	Population	Genotype call in both	% same genotype called	% different genotype called
WA_1	WA	36220	98.32	1.68
WA_2	WA	16006	98.43	1.57
WA_3	WA	33342	98.46	1.54
WA_4	WA	10576	98.08	1.92
WA_5	WA	28630	98.70	1.30
WA_6	WA	34987	98.52	1.48
WA_7	WA	15316	98.70	1.30
WA_8	WA	19201	98.90	1.10
WA_9	WA	30852	97.08	2.92
WA_10	WA	15355	98.51	1.49
WA_11	WA	7938	99.26	0.74
WA_12	WA	50728	98.09	1.91
WA_13	WA	41639	98.34	1.66
WA_14	WA	40977	98.42	1.58
WA_15	WA	14530	99.11	0.89
Mean	**WA**	**26420**	**98.46**	**1.54**
Median	**WA**	**28630**	**98.46**	**1.54**
SK_1	SK	110335	98.14	1.86
SK_2	SK	78353	98.70	1.30
SK_3	SK	107044	98.02	1.98
SK_4	SK	83367	97.98	2.02
SK_5	SK	103565	97.87	2.13
SK_6	SK	111951	98.18	1.82
SK_7	SK	99917	98.02	1.98
SK_8	SK	112028	98.32	1.68
SK_9	SK	102263	97.90	2.10
SK_10	SK	106678	97.75	2.25
SK_11	SK	103438	97.46	2.54
SK_12	SK	117917	98.59	1.41
SK_13	SK	56338	98.78	1.22
SK_14	SK	111478	98.06	1.94
SK_15	SK	106235	98.05	1.95
SK_16	SK	101067	97.96	2.04
Mean	**SK**	**100748**	**98.11**	**1.89**
Median	**SK**	**104900**	**98.04**	**1.96**
Mean Overall		**64783**	**98.28**	**1.72**
Median Overall		**56338**	**98.32**	**1.68**

Table S5. Comparison of the SNP and genotype calling of SAMtools and CLC for each individual

Individual	Population	Genotype call in both	% same genotype called	% different genotype called
WA_1	WA	37882	96.21	3.79
WA_2	WA	16683	96.26	3.74
WA_3	WA	35020	95.61	4.39
WA_4	WA	11181	95.49	4.51
WA_5	WA	29655	96.57	3.43
WA_6	WA	36447	96.13	3.87
WA_7	WA	15789	97.37	2.63
WA_8	WA	19888	96.75	3.25
WA_9	WA	31782	92.46	7.54
WA_10	WA	16227	96.09	3.91
WA_11	WA	7976	97.24	2.76
WA_12	WA	52999	95.25	4.75
WA_13	WA	43371	96.15	3.85
WA_14	WA	42776	95.71	4.29
WA_15	WA	14950	97.61	2.39
Mean	**WA**	**27508**	**96.06**	**3.94**
Median	**WA**	**29655**	**96.15**	**3.85**
SK_1	SK	108741	97.45	2.55
SK_2	SK	73521	97.82	2.18
SK_3	SK	106153	97.51	2.49
SK_4	SK	81675	97.32	2.68
SK_5	SK	102473	97.26	2.74
SK_6	SK	109460	97.51	2.49
SK_7	SK	98369	97.37	2.63
SK_8	SK	108453	97.59	2.41
SK_9	SK	100769	97.27	2.73
SK_10	SK	105498	97.20	2.80
SK_11	SK	98858	96.42	3.58
SK_12	SK	111969	97.78	2.22
SK_13	SK	53262	97.82	2.18
SK_14	SK	108561	97.18	2.82
SK_15	SK	103734	97.59	2.41
SK_16	SK	99683	97.44	2.56
Mean	**SK**	**98199**	**97.41**	**2.59**
Median	**SK**	**103104**	**97.45**	**2.55**
Mean Overall		**63994**	**96.76**	**3.24**
Median Overall		**53262**	**97.24**	**2.76**

Table S6. List of restriction enzymes

Enzyme	Sequence
AluI	AGCT
BstUI	CGCG
DpnI	GATC
HaeIII	GGCC
RsaI	GTAC
HpyCH4V, CvRI	TGCA
AfeI	AGCGCT
BmgBI	CACGTC
BsrBI	CCGCTC
BstZ17I	GTATAC
DraI	TTTAAA
Eco53kI	GAGCTC
EcoRV	GATATC
FspI	TGCGCA
HpaI	GTTAAC
MscI	TGGCCA
NaeI	GCCGGC
PvuII	CAGCTG
ScaI	AGTACT
SfoI	GGCGCC
SmaI	CCCGGG
StuI	AGGCCT
ZraI	GACGTC

Table S7. Primer sequences used for the genotype validations. Bases in small letters indicate low quality bases in the orangutan reference genome. Forward primers are indicated by '_F', reverse primers by '_R'.

Primer ID	Primer Sequence (5'-3')	Primer ID	Primer Sequence (5'-3')
chr1_24944851_F	aaacaaactcattgccgaaag	chr9_127539021_F	CCTCGTCATAGGCAAAGGTAAG
chr1_219461506_F	CAATCCTTGGCCTGATGAA	chr9_124746781_F	CGAAGTGTGAAGCACCaactaa
chr1_175192409_F	caccacacctggcaatttt	chr9_47115306_F	GATGATGATGTCCTGCTCCA
chr1_208620414_F	CCAGTGTGTGTGAAGCAAATTC	chr9_134266946_F	GCCGAGAGCAACATGAATGA
chr1_107518541_F	CTAGGAAAGAGCACATTGGGAAG	chr9_41701746_F	GGCCATGCCAGATGAGG
chr1_185953605_F	CTGAACCTCCAGCATCCAAC	chr9_23407926_F	TCAGTTCCTCCCCACCATT
chr1_4242060_F	gactatggaagtcataagaagcgagt	chr9_90401891_F	ttatgatttgtccccgttactctatc
chr1_112009373_F	GCCACCCAGATGACAGCA	chr9_124605789_F	TTCCAGCAGTCTTGGGTCAG
chr1_5139357_F	GCTCGAATGCTTGTGTGAGG	chr10_65000478_F	AAACTGTATCTGGGAAAGGATGA
chr1_226543536_F	GGAAATGGCCTCAAGGAAGTA	chr10_103756133_F	AAGAGGTGAGGGGCAGGT
chr1_143391350_F	ggcctctcttcttggcttg	chr10_3235539_F	ACAGGATGCTCGCCATCTAT
chr1_3958934_F	GTCTCTGTAGCTCGGCTTCC	chr10_108961144_F	AGACAAGGGAGGTGTGAAAAC
chr1_221293555_F	TCTCCTAGCGGCACCTGA	chr10_130576144_F	CATTTGAGTGTAAGGGTGGGTTTA
chr2a_28293664_F	acctctccatcttcttagggtcagt	chr10_103414836_F	CCAAACCAAACAGCCTTCC
chr2a_33857975_F	AGGCAGGCATAGGGAATTAG	chr10_133022276_F	ccgcctcggctacctaa
chr2a_6106081_F	gcagctagggtggtaaacAGAG	chr10_5043197_F	tcccacatcccgagtcc
chr2a_110169000_F	TGGAGCTATCACAGGACGATG	chr11_130432197_F	aggcatggtgctaggaacttaac
chr2b_116732260_F	ccaagacagcgagagagagagag	chr11_45914047_F	catgctctgcacctacatttctttt
chr2b_24768739_F	CTGCGCTTTTAACCTCATACAC	chr11_4072568_F	GGAGCTAGAGAGCCAGAAAGA
chr2b_64996248_F	GGAGAAGGAAAGAACCCACTTA	chr11_8660875_F	GGGGAATGGCGAAGGTT
chr2b_133348340_F	GGCCGGTGTTTCGAGAG	chr11_117174368_F	TCCAGGCTCAACTTTGTTCAC
chr2b_6245962_F	tctggctttaggggatcttgtt	chr12_51604109_F	AACGAGAGAGACCTAATTGGCTAC
chr2b_101269295_F	TGTTCTGTAAACACCACCTAATTG	chr12_1624763_F	GGCCTCGCGTTTGGTAg
chr3_4892749_F	ACATGACCTTTAGTGGGCAAAG	chr12_239824_F	TTCCCAGGCCAACCTTTAC
chr3_79444560_F	ACTTTGCTCTAAAACCATTTGTGTC	chr12_126793776_F	TTCCTGTTGAAGCCCATCTG
chr3_158213691_F	gattacaggcgtgagcGTTA	chr13_30715519_F	GGTTCTCAGGTTTAGAGGGTGAT
chr3_8251414_F	GCTGAGGTCATACTTGGTAAAGGAG	chr14_96790956_F	AGAGGGTAAAAGGTAGGGGTGA
chr3_135779953_F	GGCCCTTTGCCTATTTGTC	chr14_101297454_F	ctccccatatccacccactt
chr4_82339163_F	GAGGGCAAAGGCTGAACTTG	chr14_106505114_F	GGTGGCATAAAGACCAAAAGGT
chr5_150282746_F	ATCCAGCATAAAGAACcctctctg	chr14_79127075_F	TTGGGAGTTCACCTACAATGC
chr5_15192012_F	CAGCAGGCGTTGGACAG	chr15_87539423_F	gccaccacgcctaatttt
chr5_183803425_F	GAACGTGTACTCAAATCCTCCTCT	chr16_71973781_F	accctccatccctgtttaatttt
chr5_12085608_F	GAACTAAATGGTGGGTTGAGTGT	chr16_54132375_F	CACAAGGTCAGCCAAGAGC
chr5_175571985_F	GAGTCCTCGCCCTCCTTAGC	chr16_73068101_F	TGGAAGAAGGGGCGTGT
chr6_32510526_F	AACTCTGTCCCTGGAATTGAAA	chr17_9583225_F	AATTTGTGGCTGCATGAGG
chr6_32510526_F	AACTCTGTCCCTGGAATTGAAA	chr17_64992963_F	CCAGACGAGAGACTGAATAAAGAG
chr6_70010056_F	ACTGATGAACCCTCGTCTGTG	chr17_29515153_F	GAAAACCGTGATATGGCTCAC
chr6_35351998_F	CATCCCAAAGGGCCAAG	chr17_43041507_F	GTTCTCTGAAGACCGGAACC
chr6_30843959_F	CCTGGGTCCTCCCTGATAG	chr17_14277696_F	TGTGTTGTGAAGTAAAAGCTGGAA
chr6_7054385_F	GTTATGACCCAGATACGTGGTG	chr18_19718397_F	GGGAACTCCTCACGGATCTT
chr6_25321401_F	TCTAAATACCAACACTTAACCCAGA	chr19_2253926_F	ACATGCCCAAAATCACTGG
chr7_11255797_F	CATGACGTTTGTAAATGCCTCTAGT	chr19_6967639_F	GTCCTTGGTGTCTCTTTGCAG
chr7_28221055_F	CCTTTCCATCGTGCTGGT	chr20_29197787_F	atagagtttgttgggagaagtgg
chr7_148478776_F	CTCCAATGAAATCTGCGAAAA	chr20_61451986_F	CATCCCACTGACCCGAAAT
chr7_33107094_F	gatggggttttgccatgttg	chr20_62514743_F	CCCACCGGGCCTTAGTT
chr7_32762731_F	TATGTCCCTGTGAAGTGTTAAGAGA	chr20_42404631_F	cctcctgagttcaaacgattctc
chr7_32727906_F	TTTTGAAGAAGACAGAAGGATGG	chr20_35898880_F	GGATGTAAGCccggatctgt
chr8_148777043_F	TGCCAAACTGTGCTCCCTAT	chr21_13339641_F	GTAGAAATTGGCCTTTTGGACT
chr22_18310003_F	CCCCGTTTGGAAGGAGTAGAG	chr22_15595242_F	GGCTTTGACTTCACGCTCAT
chr22_31379523_F	CGGTGGAAAGAATGCTCAC	chrX_149542552_F	ACAGGCATGGTTCAAGTTCC

Table S7 (Continued)

chrX_66733001_F	AGCTGAGAACACATTCCCTGT	chr9_47115306_R	GATGATGATGTCCTGCTCCA
chrX_119397713_F	cccagttccacctttaacacc	chr9_134266946_R	CCTCCCTACCCCTTTTGG
chrX_10406492_F	TGTGTTGGCCTGGGTATGAC	chr9_41701746_R	GGCCATGCCAGATGAGG
chr1_24944851_R	TGGTGTGTGGGCCTAGC	chr9_23407926_R	gctgaagatacagttggcagtctt
chr1_219461506_R	CAATCCTTGGCCTGATGAA	chr9_90401891_R	AGCCTGGAAACGCCATCTA
chr1_175192409_R	ccctaataacccctgcctct	chr9_124605789_R	ATGGCTCGTGTGGTTTGG
chr1_208620414_R	CCAGTGTGTGTGAAGCAAATTC	chr10_65000478_R	AAACTGTATCTGGGAAAGGATGA
chr1_107518541_R	CAGACTCACCTGCGACCTGT	chr10_103756133_R	TCTCTCGTATTGCCTACAAAATG
chr1_185953605_R	CTGAACCTCCAGCATCCAAC	chr10_3235539_R	TTGTATTTCAGTGGGTGCTTGT
chr1_4242060_R	ccatctcttggccttactcaatc	chr10_108961144_R	AGACAAGGGAGGTGTGAAAAC
chr1_112009373_R	TCACTCCGCAGGCCAGA	chr10_130576144_R	ATGAGACTTGCGGCTTGG
chr1_5139357_R	GGTGCAGCCCACGTTGT	chr10_103414836_R	CCAAACCAAACAGCCTTCC
chr1_226543536_R	gcagggacacagtttagcc	chr10_133022276_R	GGTGTCATGTCCCGTGCTAA
chr1_143391350_R	gcttgtgccatccccatt	chr10_5043197_R	tgcaaaggcacagaggtg
chr1_3958934_R	GTCTCTGTAGCTCGGCTTCC	chr11_130432197_R	AATCCTGTGCTCCCAGAAGTG
chr1_221293555_R	TCTCCTAGCGGCACCTGA	chr11_45914047_R	TGTTGCTCTTCTAGGCAGTGTTCT
chr2a_28293664_R	ggcctaccatagcaaaactcaaa	chr11_4072568_R	GGAGCTAGAGAGCCAGAAAGA
chr2a_33857975_R	AGGCAGGCATAGGGAATTAG	chr11_8660875_R	GGGGAATGGCGAAGGTT
chr2a_6106081_R	gcagctagggtggtaaacAGAG	chr11_117174368_R	tggcattacaggcgtgagt
chr2a_110169000_R	GCTGTCAAGGGGCGAAA	chr12_51604109_R	AACGAGAGAGACCTAATTGGCTAC
chr2b_116732260_R	caaagtgccataaactgggtagc	chr12_1624763_R	GGCCTCGCGTTTGGTAg
chr2b_24768739_R	CCAGGCTTCGCTCTTTG	chr12_239824_R	TTCCCAGGCCAACCTTTAC
chr2b_64996248_R	GGAGAAGGAAAGAACCCACTTA	chr12_126793776_R	TTGGCCTGGAAAATTAGTTGCT
chr2b_133348340_R	TGACAGTGAGGAAGAAGGGTAAA	chr13_30715519_R	GGTTCTCAGGTTTAGAGGGTGAT
chr2b_6245962_R	gaggcttgggctgattttct	chr14_96790956_R	CGGATGCCAGAGGAATTG
chr2b_101269295_R	AAAACATCTGGGGAGGCTATAA	chr14_101297454_R	CCACTCTCAACTTTCATTCTCATTC
chr3_4892749_R	ACATGACCTTTAGTGGGCAAAG	chr14_106505114_R	GGTGGCATAAAGACCAAAAGGT
chr3_79444560_R	ACTTTGCTCTAAAACCATTTGTGTC	chr14_79127075_R	TTGGGAGTTCACCTACAATGC
chr3_158213691_R	gattacaggcgtgagcGTTA	chr15_87539423_R	gccaccacgcctaatttt
chr3_8251414_R	GGATTTTGCTCGACAGTGCAT	chr16_71973781_R	cattttagacgctggggttacag
chr3_135779953_R	GGCCCTTTGCCTATTTGTC	chr16_54132375_R	CCTCCCCATAACATTCAGTACAC
chr4_82339163_R	GAATATGAAGAtgtggtgctggact	chr16_73068101_R	ctggaaagtggagacattagtcc
chr5_150282746_R	GGTGTCTGCAAAGAAACACGGTA	chr17_9583225_R	AATTTGTGGCTGCATGAGG
chr5_15192012_R	CAGCAGGCGTTGGACAG	chr17_64992963_R	CCAGACGAGAGACTGAATAAAGAG
chr5_183803425_R	GAACGTGTACTCAAATCCTCCTCT	chr17_29515153_R	GAAAACCGTGATATGGCTCAC
chr5_12085608_R	GAACTAAATGGTGGGTTGAGTGT	chr17_43041507_R	GTTCTCTGAAGACCGGAACC
chr5_175571985_R	AAGCAGACGATAAACTGCAATCC	chr17_14277696_R	ttgtgcccatgtaccctagaa
chr6_32510526_R	AACTCTGTCCCTGGAATTGAAA	chr18_19718397_R	GGGAACTCCTCACGGATCTT
chr6_32510526_R	TGTCGGAGCCAAGTAGATCA	chr19_2253926_R	ACATGCCCAAAATCACTGG
chr6_70010056_R	ACTGATGAACCCTCGTCTGTG	chr19_6967639_R	GTCCTTGGTGTCTCTTTGCAG
chr6_35351998_R	CCCACCAGGGTAGTTATCAATC	chr20_29197787_R	atagagtttgttgggagaagtgg
chr6_30843959_R	CCTGGGTCCTCCCTGATAG	chr20_61451986_R	GACAGGCAGAGGACAAACG
chr6_7054385_R	GTTATGACCCAGATACGTGGTG	chr20_62514743_R	CCCACCGGGCCTTAGTT
chr6_25321401_R	TCTAAATACCAACACTTAACCCAGA	chr20_42404631_R	TGGCCGCTGATCCTCAA
chr7_11255797_R	GGAAGTAGAACTGGGCAAGAAAG	chr20_35898880_R	CCTCACTCCAACTGGGTCTTT
chr7_28221055_R	CCTTTCCATCGTGCTGGT	chr21_13339641_R	GTAGAAATTGGCCTTTTGGACT
chr7_148478776_R	CTCCAATGAAATCTGCGAAAA	chr22_18310003_R	GAGTACCCGCAGGAGGACA
chr7_33107094_R	cagcattcggtgtctggtga	chr22_31379523_R	CGGTGGAAAGAATGCTCAC
chr7_32762731_R	TATGTCCCTGTGAAGTGTTAAGAGA	chr22_15595242_R	CCACTCGGTGTTGTTGACAG
chr7_32727906_R	TTTTGAAGAAGACAGAAGGATGG	chrX_149542552_R	ACAGGCATGGTTCAAGTTCC
chr8_148777043_R	TGCCAAACTGTGCTCCCTAT	chrX_66733001_R	AGCTGAGAACACATTCCCTGT
chr9_127539021_R	CCTCGTCATAGGCAAAGGTAAG	chrX_119397713_R	gatttctttggctgtaagtaacgtct
chr9_124746781_R	CGAAGTGTGAAGCACCaactaa	chrX_10406492_R	TGTGTTGGCCTGGGTATGAC

Chapter 4

Orangutan demographic history and population structure inferred by genus-wide whole-genome sequencing

Maja P. Greminger[1], Alexander Nater[2,1], Javier Prado-Martinez[3], Benoit Goossens[4,5,6], Ernst Verschoor[7], Kristin Warren[8], Ian Singleton[9,10], Ivo Gut[11], Marta Gut[11], Laurentius N. Ambu[6], Carel P. van Schaik[1], Tomas Marques-Bonet[3,11], and Michael Krützen[1]

[1]Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

[2]Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

[3]CREA, Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain

[4]Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom

[5]Danau Girang Field Centre, c/o Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[6]Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[7]Biomedical Primate Research Centre, Rijswijk, The Netherlands

[8]School of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, Australia

[9]Foundation for a Sustainable Ecosystem (YEL), Medan, Indonesia

[10]PanEco, Foundation f or Sustainable Development and Intercultural Exchange, Berg am Irchel, Switzerland

[11]Centro Nacional de Análisis Genómico, Barcelona, Spain

4.1 Abstract

Unraveling how different evolutionary processes have shaped patterns of genetic diversity within and among species is a main interest of evolutionary genetics. To disentangle neutral stochastic variation from the effects of selection, detailed knowledge of a species' demographic history and population structure is required. Orangutans, currently endemic to the Sundaland islands of Borneo (*Pongo pygmaeus*) and Sumatra (*Pongo abelii*), have likely experienced a complex demographic history. Southeast Asian Sundaland has been drastically affected by Pleistocene climatic oscillations, sea level changes, and volcanic activities. We investigated the demographic history and population structure of the genus *Pongo* by applying the most comprehensive genomic sampling to date, encompassing whole genomes of 36 orangutans representing the entire extant geographic distribution of the genus. We found that the speciation of Bornean and Sumatran orangutans has been a long-lasting gradual process, strongly impacted by recurrent climate changes. Following their initial separation in the early Pleistocene, autosomal gene pools remained connected via regular gene flow over the cyclically exposed Sunda Shelf during glacial periods. Autosomal gene pools appear to have started diverging ~0.9–1.1 Ma, indicating a substantial reduction of gene flow levels. Evolutionary trajectories of Bornean and Sumatran orangutans differed drastically subsequently. Bornean orangutans likely experienced several bottlenecks and a long-term population decline, most probably related to climate and thus habitat fluctuations. In contrast, Sumatran orangutans exhibited a remarkably stable population history throughout the Pleistocene and seem to have been much less affected by climate oscillations, likely due to the different geology and environmental conditions on Sumatra. We inferred, however, that the population size of Sumatran orangutans collapsed drastically coinciding with the Toba supereruption ~73 ka. To our knowledge, this represents the first genetic evidence of a strong regional impact of the supereruption on a large mammal. The autosomal genome data further confirmed pronounced extant population structure on both islands. Most strikingly, Batang Toru, the only remaining population south of Lake Toba, was clearly distinct from all other Sumatran orangutans, probably caused by the recurrent activity of the Toba volcano. This finding is congruent with previous results from mitochondrial DNA and small-scale autosomal markers. Given the genetic uniqueness of the orangutans south of Lake Toba, we propose a taxonomic revision of *P. abelii* as well as to manage Batang Toru as a separate evolutionary significant unit. Since the census size of the Batang Toru population is already reduced to few hundreds individuals, urgent conservation efforts are required.

4.2 Introduction

Understanding evolutionary forces that shaped patterns of genetic diversity has been a long-standing goal of evolutionary biology (e.g. Hahn 2008). Such patterns within and among species are the result of demography, selection and stochasticity. Estimating the relative importance of these different processes is challenging (e.g. Nei *et al.* 2010). Under certain demographic events, such as population subdivision or population size changes, random genetic drift can lead to similar signals in the genome as natural selection (Tajima 1989; Andolfatto & Przeworski 2000; Nielsen 2005; Teshima *et al.* 2006; Excoffier *et al.* 2009). Therefore, detailed knowledge of the demographic history and population structure is required to unravel their confounding effects from signals of selection (Wall *et al.* 2002; Haddrill *et al.* 2005; Nielsen *et al.* 2005b; Stajich & Hahn 2005).

Orangutans (genus: *Pongo*), the only Asian great apes, show remarkable geographic variation in various functional traits (van Schaik *et al.* 2009b; Wich *et al.* 2009b), suggesting high levels of local adaptions. Due to the basal position of the genus *Pongo* in the great ape lineage, the genus is of high interest when reconstructing of the adaptive evolutionary history of great apes. Orangutans, however, have experienced an eventful demographic history with major changes in their distribution during the Pleistocene (von Koenigswald 1982; Rijksen & Meijaard 1999; Delgado & van Schaik 2000). They were once widely distributed throughout mainland Southeast Asia and most of the Sundaland islands (von Koenigswald 1982; Rijksen & Meijaard 1999; Delgado & van Schaik 2000). Their current range, however, is restricted to increasingly isolated forest patches in northern Sumatra (*P. abelii*), while they show a wider distribution on Borneo (*P. pygmaeus*) (Figure 1; Wich *et al.* 2008).

The Sunda archipelago has been drastically affected by geological and environmental processes such as tectonic plate movements, Quaternary climatic oscillations, fluctuating sea levels, and volcanic eruptions (Hall 2002; Bird *et al.* 2005). Since the Sundaland islands attained their present shape in the Early Pleistocene (Meijaard 2004), the continental shelf was cyclically exposed during glacial periods when sea levels were lower (Verstappen 1997; Voris 2000), potentially admitting terrestrial migration between islands. The Sunda archipelago was also impacted by a series of strong volcanic eruptions, mainly on Sumatra and Java (Hall 1996). Most notable is Mount Toba on northern Sumatra, which had at least four major and various smaller eruptions during the last 1.2 million years (Hall 1996; Williams *et al.* 2009). The Toba supereruption ~73 ka is seen as the largest eruption of the Quaternary (Chesner *et al.* 1991). Nevertheless, the consequences of this supereruption on regional and global climate, terrestrial ecosystems, and prehistoric human populations remain highly controversial (e.g. Haslam & Petraglia 2010; Williams *et al.* 2010; Williams 2012).

The evolutionary history of orangutans has been strongly influenced by the aforementioned environmental factors (e.g. Delgado & van Schaik 2000; Warren *et al.* 2001; Steiper 2006; Arora *et al.* 2010; 2011; Nater *et al.* 2013; Nater *et al.* 2015). For example, despite current census size of Sumatran orangutans being almost tenfold smaller than that of Bornean

orangutans (Wich *et al.* 2008), genetic diversity and long-term effective population size (N_e) are much higher than for Bornean orangutans (Steiper 2006; Locke *et al.* 2011; Nater *et al.* 2011; Prado-Martinez *et al.* 2013). This pattern has been found for both autosomal and mitochondrial sequences. The lower genetic diversity in Bornean orangutans may be linked to a severe bottleneck during the penultimate glacial period (190–130 ka), during which Bornean orangutans might have been forced into a common rainforest refugium (Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2015).

Extant orangutans exhibit complex population structure as documented in genetic studies of both biparentally inherited microsatellite markers and mitochondrial loci (Chapter 3; Warren *et al.* 2001; Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015). The major genetic clusters of orangutans—hereafter referred to as populations—are separated by barriers such as large rivers or mountain ridges (Figure 1). The particularly complex genetic structure of Sumatran orangutans is highlighted by the lack of reciprocal monophyly for mitochondrial gene sequences between both currently recognized species in the genus *Pongo*, where the lineage of Batang Toru—the only remaining Sumatran population south of the Toba caldera—is more closely related to the lineage leading to Bornean orangutans than to other Sumatran orangutans (Nater *et al.* 2011).

There is strong incongruence among divergence time estimates between Bornean and Sumatran orangutans, probably due to the different genetic marker systems that were employed. Inferences from mitochondrial loci ranged from a split 1 to 5 million years ago (Ma) (Zhi *et al.* 1996; Warren *et al.* 2001; Zhang *et al.* 2001; Steiper 2006; Nater *et al.* 2011). Recent analyses of the autosomal genome, however, suggest a much more recent divergence ~330–600 ka (Locke *et al.* 2011; Mailund *et al.* 2011; Mailund *et al.* 2012). Using a few Y-linked markers, Nater *et al.* (2011) found an unexpectedly recent coalescence for Bornean and Sumatran orangutans of only 168 ka, suggesting recent gene flow between islands. Yet, it remains highly controversial to what extent the recurrent land bridges between Borneo and Sumatra during Pleistocene glacial periods allowed for migration of orangutans and other rainforest-dependent species between islands (Gorog *et al.* 2004; Harrison *et al.* 2006; Kanthaswamy *et al.* 2006; Steiper 2006; Nater *et al.* 2011; Nater *et al.* 2015).

Although previous genetic studies offered important insights into the complex evolutionary history and phylogeography of orangutans, genetic diversity present in the autosomal genome remains poorly studied. Current autosomal data of wild orangutans with known population provenance are largely restricted to short DNA sequences and few microsatellite markers genotyped in non-invasively collected samples (Chapter 3; Kanthaswamy *et al.* 2006; Arora *et al.* 2010; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015). Recently, Locke *et al.* (2011) and Prado-Martinez *et al.* (2013) deep-sequenced whole-genomes of ten captive orangutans. However, despite most individuals were wild-caught, the actual provenance of these individuals were unknown. This obstacle, together with their small sample sizes, limited the conclusions that could be drawn regarding the extent and distribution of genetic diversity in wild populations.

Here, we genetically assigned the previously sequenced orangutans (Locke *et al.* 2011; Prado-Martinez *et al.* 2013) to their natal populations based on our detailed knowledge of orangutan phylogeography and population structure (Chapter 3; Arora *et al.* 2010; Nater *et al.* 2011; Nietlisbach *et al.* 2012; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015), which provides a *hitherto* unprecedented opportunity to attain sample provenance retrospectively. We also complemented previous sequencing efforts by re-sequencing genomes of eleven wild-born orangutans to medium–high coverage. In total, our sample set comprised genomes of 36 unrelated orangutans, representing the entire current geographic range of the genus *Pongo* (Figure 1, Table 1). We used these genomes to study the geographic structure of autosomal genetic diversity in orangutans. Furthermore, we investigated the demographic history of the genus *Pongo* in light of the complex environmental forces of Sundaland.

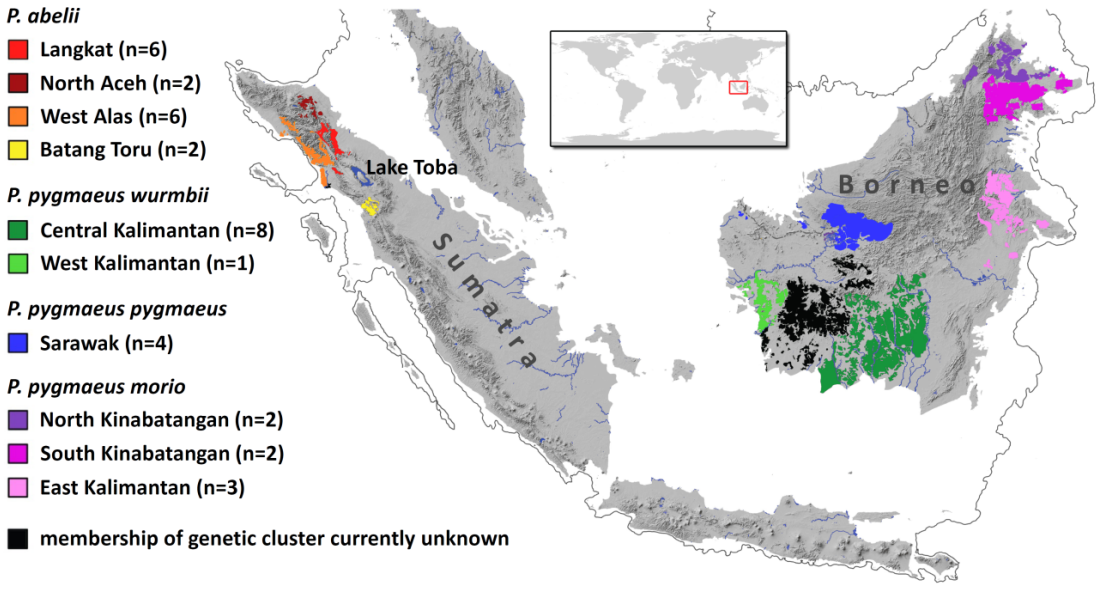


Figure 1. Geographic distribution of the genus *Pongo*. Orangutan sampling areas are represented by different colors. Samples sizes for each population are given in parentheses. The grey line around the islands indicates the extent of the exposed Sunda shelf during the last glacial maximum (~120 meters below current sea level). Extant orangutan sampling areas are: Langkat (LK, east of Alas River, south of Tamiang River), North Aceh (NA, east of Alas River, north of Tamiang River), West Alas (WA, west side of Alas River, including Tripa, West Leuser, Central Leuser and Batu Ardan), Batang Toru (BT, south of Lake Toba but north of the Batang Toru river), South Kinabatangan (SK, south of the Kinabatangan River), North Kinabatangan (NK, north of the Kinabatangan River), East Kalimantan (EK, north of the Mahakam River and south of the Kayan River), Sarawak (SR, north of the Kapuas river), Central Kalimantan (CK), and West Kalimantan (WK).

4.3 Materials and Methods

Sampling schema and population assignment

Whole-genome sequence data for the 36 study individuals were obtained from four different sources (Table 1, Supporting Table S1): (i) genomes of eleven orangutans were sequenced in this study. Data for 20 individuals were obtained from (ii) Locke *et al.* (2011, n=10) and (iii) Prado-Martinez *et al.* (2013, n=10). (iv) We supplemented our dataset with five unpublished orangutan genomes previously sequenced by Prado-Martinez *et al.* (2013). All individuals were wild-born, except of five orangutans which were first-generation offspring of wild-born parents of the same species (Supporting Table S1).

We identified the most likely natal area of study individuals based on mtDNA haplotype clustering in a phylogenetic tree together with samples of known geographic provenance. Because of strict female philopatry in orangutans, mtDNA haplotypes are reliable indicators for the population of origin (Arora *et al.* 2010; Morrogh-Bernard *et al.* 2011; Nater *et al.* 2011; Arora *et al.* 2012; Nietlisbach *et al.* 2012; van Noordwijk *et al.* 2012). Using three concatenated mtDNA genes (16S ribosomal DNA, Cytochrome b, and NADH-ubiquinone oxidoreductase chain 3), we constructed a Bayesian tree, including 127 non-invasively sampled wild orangutans from 15 geographic regions representing all known extant orangutan populations (Nater *et al.* 2011; Nater *et al.* 2015). Gene sequences of our study individuals were extracted from their complete mitochondrial genome sequences (cf. Chapter 5). The phylogenetic tree was built with BEAST v1.8.0. (Drummond *et al.* 2012) as described in Nater *et al.* (2011), applying a TN93+I substitution model (Tamura & Nei 1993) determined by jModelTest v2.1.4. (Darriba *et al.* 2012).

We were able to assign all previously sequenced orangutans (Locke *et al.* 2011; Prado-Martinez *et al.* 2013) to their most likely population of origin (Supporting Table S1, Supporting Figures S1–S3). The sample assignment revealed incomplete geographic representation of the genus *Pongo* by previous sequencing studies (Table 1 and Supporting Table S1): orangutans sequenced by Locke *et al.* (2011) were assigned to three of the six Bornean populations, and to three of the four Sumatran populations. Individuals sequenced by Prado-Martinez *et al.* (2013) were assigned to two Bornean and two Sumatran populations, respectively. Thus, in order to achieve more complete representation of extant orangutans, we sequenced genomes of eleven wild-born orangutans (Figure 1, Supporting Table S1) mainly from areas others than covered before. Detailed provenance information for these individuals is provided in Supporting Table S1.

Whole-genome sequencing

To obtain sufficient amounts of DNA, we collected blood samples from confiscated orangutans at rehabilitation centers, including the Sumatran Orangutan Conservation Program (SOCP) in Medan, BOS Wanariset Orangutan Reintroduction Project in East

Kalimantan, Semongok Wildlife Rehabilitation Centre in Sarawak, and Sepilok Orangutan Rehabilitation Centre in Sabah. Whole blood samples were taken during routine veterinary examinations and stored in EDTA blood collection tubes at -20°C. The collection and transport of samples were conducted in strict accordance with Indonesian, Malaysian and international regulations. Samples were transferred to Zurich under the Convention on International Trade of Endangered Species in Fauna and Flora (CITES) permit numbers 4872/2010 (Sabah), and 06968/IV/SATS-LN/2005 (Indonesia), respectively.

Genomic DNA was extracted with the Gentra Puregene Kit (Qiagen) using a modified protocol for clotted blood as described in Chapter 3 (Greminger *et al.* 2014). Individuals were sequenced on two to three lanes on a Illumina HiSeq 2000 in paired end (2 x 101 bp) mode. Sample PP_5062 was sequenced at the Functional Genomics Center in Zurich (Switzerland), the other individuals at the Centre Nacional d'Anàlisi Genòmica in Barcelona (Spain) like the individuals of Prado-Martinez *et al.* (2013).

Table 1. Overview sampling orangutan whole-genome sequencing. Effective read-depths (see main text) are given as ranges in brackets below the source reference. Sample sizes in parenthesis correspond to captive-born individuals.

				This study		
		Locke *et al.* 2011	**Prado-Martinez *et al.* 2013**	**PM unpubl.[a]**	**Novel**	
Species	**Sampling areas**	[4.8-12.2x]	[20.5-27.4x]	[11.1-25.3x]	[13.7-31.1x]	**Total**
P. abelii	Langkat (LK)	2	4	0	0	6
P. abelii	North Aceh (NA)	0	(1)	0	1	1+(1)
P. abelii	West Alas (WA)	2	0	2	2	6
P. abelii	Batang Toru (BT)	1	0	0	1	2
P. pygmaeus	South Kinabatangan (SK)	0	0	0	2	2
P. pygmaeus	North Kinabatangan (NK)	0	0	0	2	2
P. pygmaeus	East Kalimantan (EK)	1	0	0	2	3
P. pygmaeus	Sarawak (SR)	1	1	1+(1)	0	3+(1)
P. pygmaeus	Central Kalimantan (CK)	3	2+(2)	(1)	0	5+(3)
P. pygmaeus	West Kalimantan (WK)	0	0	0	1	1

[a]unpublished genomes sequenced by Prado-Martinez *et al.* (2013)

Read mapping, SNP and genotype calling

We followed identical bioinformatical procedures for all 36 study individuals using same software versions.

Read mapping

Raw Illumina sequencing reads were quality-checked with FastQC v0.10.1. (Andrews 2012) and mapped to the orangutan reference genome *PonAbe2* (Locke *et al.* 2011) using the Burrows-Wheeler Aligner (BWA-MEM) v0.7.5 (Li & Durbin 2009) in paired-end mode with

default read alignment penalty scores. Picard v1.101 (http://picard.sourceforge.net/) was used to add read groups, convert sequence alignment/map (SAM) files to binary alignment/map (BAM) files, merge BAM files for each individual, and mark optical and PCR duplicates. We filtered out duplicated reads, bad read mates, reads with mapping quality zero, and reads which mapped ambiguously.

We performed local realignment around indels and empirical base quality score recalibration (BQSR) with the Genome Analysis Toolkit (GATK) v3.2.2. (McKenna *et al.* 2010; DePristo *et al.* 2011). The BQSR process empirically calculates more accurate base quality scores (i.e. Phred-scaled probability of error) than those emitted by the sequencing machines through analyzing the covariation among several characteristics of a base (e.g. position within the read, sequencing cycle, previous base, etc.) and its status of matching the reference sequence or not. To account for true sequence variation in the data set, the model requires a database of known polymorphic sites ("known sites") which are skipped over in the recalibration algorithm. Since no suitable set of "known sites" was available for the complete genus *Pongo*, we preliminary identified confident SNPs from our data. For this, we performed an initial round of SNP calling on unrecalibrated BAM files with the *Unified Genotyper* of the GATK. SNPs were called separately for each orangutan species in multi-sample mode (i.e. joint analysis of all individuals per species), creating two variant call (VCF) files. In addition, we produced a third VCF file jointly analyzing all study individuals in order to capture genus-wide low frequency alleles. We applied the following hard quality filter criteria on all three VCF files: 'QUAL < 50.0 || QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0'. Additionally, we applied a custom-made perl script to determine mean and standard deviation of sequencing depth over all samples and filtered all sites with a site-wise coverage more than five standard deviations above the mean. The three hard filtered VCF files were merged and SNPs taken as "known sites" for BQSR with the GATK. The walkers *CountReads* and *DepthOfCoverage* of the GATK were used to obtain various mapping statistics for unfiltered and filtered BAM files.

SNP and genotype calling

We identified SNPs and called genotypes in a three-step approach. First, we identified a set of candidate (raw) SNPs among all study individuals. Second, we performed variant quality score recalibration (VQSR) on the candidate SNPs to identify high-confidence SNPs. Third, we called genotypes of all study individuals at these high-confidence SNP positions.

Step 1: we used the *HaplotypeCaller* of the GATK in genomic VCF (gVCF) mode to obtain for each individual in the dataset genotype likelihoods at any site in the reference genome. *HaplotypeCaller* performs local realignment of reads around potential variant sites and is therefore expected to considerably improve SNP calling in difficult-to-align regions of the genome. The resulting gVCF files were then genotyped together on a per-species level, as well as combined for all individuals in both species, using the *Genotype GVCFs* tool of the

GATK to obtain three VCF files with candidate SNPs for *P. abelii*, *P. pygmaeus,* and over all *Pongo* samples.

Step 2: of the produced set of candidate SNPs, we identified high-confidence SNPs using the VQSR procedure implemented in the GATK. The principle of the method is to develop an estimate of the relationship between various SNP call annotations (e.g. total depth, mapping quality, strand bias, etc.) and the probability that a SNP is a true genetic variant. The model is determined adaptively based on a set of "true SNPs" (i.e. known variants) provided as input. Our "true SNPs" set contained 5,600 high-confidence SNPs, which were independently identified by three different variant callers in a previous reduced-representation sequencing project (Chapter 3; Greminger *et al.* 2014). We run the *Variant Recalibrator* of the GATK separately for each of the three raw SNP VCFs to produce recalibration files based on the "true SNPs" and a VQSR training set of SNPs. The VQSR training sets were derived separately for each of the three raw SNP VCF files and contained the top 20% SNPs with highest variant quality score after having applied hard quality filtering as described for the VCF files in the BQSR procedure. The parameters for VQSR were the following:

-resource:highconf,known=false,training=true,truth=true,prior=12.0 'true SNPs'
-resource:highconf,known=false,training=true,truth=false,prior=10.0 'training set'
-minNumBad 1000 -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an DP
-mode SNP

We used the produced VQSR recalibration files to filter the three candidate SNP VCFs with the *Apply Recalibration* walker of the GATK setting the '--truth_sensitivity _filter_level' to 99.8%. Finally, we combined all SNPs of the three VCF files passing this filter using the *Combine Variants* tool of the GATK, hence generating a master list of high-confidence SNP sites in the genus *Pongo*.

Step 3: we called the genotype of each study individual at the identified high-confidence SNP sites using the *Unified Genotyper* of the GATK applying the following command line:

"GenomeAnalysisTK.jar -T UnifiedGenotyper -R PonAbe2 -I BAM -o VCF -glm SNP -gt_mode GENOTYPE_GIVEN_ALLELES -alleles masterlist.vcf -out_mode EMIT_ALL_SITES"

Genotyping was performed on the recalibrated BAM files in multi-sample mode for Bornean and Sumatran orangutans separately, producing one SNP VCF file per species.

Finally, we only retained positions with high genome mappability, i.e. genomic positions within a uniquely mappable 100-mers (up to 4 mismatches allowed), as identified with the GEM-mappability module from the GEM library build (Derrien *et al.* 2012). This mappability mask excludes genomic regions in the orangutan reference genome that are duplicated and therefore tend to produce ambiguous mappings, which can lead to unreliable genotype calling.

Genomic consensus FASTA sequences

We also produced high-quality genomic consensus FASTA sequences for each study individual as required for the demographic history analyses. We used custom Perl scripts to create the consensus sequences by merging the information of the SNP VCF and gVCF files as following: all sites (variant and reference sites) had to be covered by at least eight individuals per species or genotypes in all individuals at this site were set to 'N'. SNP and genotype calling at genomic positions covered by fewer individuals is less accurate, hence the power to discriminate between variant and non-variant genomic positions is reduced. In addition, we filtered out sites with a mean mapping quality below 20. On the individuals' level, we required genotypes of both variant and reference positions to be covered by at least three reads, otherwise individual genotypes at that site were set to 'N'. Positions not sequenced for a given individual were also denoted as 'N'. The sequence depth of all non-variant sites in the reference genome (i.e. 'reference sites') for each individual was obtained from the gVCF files (also mappability masked) produced in the first step of the SNP calling pipeline. The genotype at variant genomic sites was extracted for each individual from the SNP VCF file described above. Heterozygous genotypes were encoded with their respective IUPAC codes.

Genetic diversity and population structure

We identified patterns of population structure in the autosomal genome by principal component analysis (PCA) of biallelic SNPs. Three separate analyses were performed: one within each species and one including all study individuals. For each sample set, we excluded all genotypes from the SNP VCF files that were covered by less than five reads and only retained SNPs with a genotype call in all individuals after this filter. Furthermore, we removed SNPs with more than two alleles and SNPs being monomorphic it the particular sample set. This restrictive filtering left us with 3,006,895 SNPs for the analysis of all study individuals, 5,838,796 SNPs for PCA within Bornean orangutans and 4,808,077 SNPs for PCA within Sumatran orangutans. The input genotype matrix files were generated using a custom Perl script and PCA performed with the R package 'prcomp'.

To further assess genetic diversity present within the identified populations, we calculated several summary statistics using custom Perl scripts. We measured heterozygosity for each individual as the mean number of heterozygous genotypes divided by the total number of called genome positions in this individual. Heterozygosity values were subsequently averaged over all individuals per population. Furthermore, we calculated nucleotide diversity (π) as the mean number of pairwise differences per site between two individuals of a given population, as well as Watterson's theta (θ_W) from the number of segregating sites. The long-term effective population sizes were estimated as $N_e = \theta/4\mu$, using a mutation rate (μ) of 1.5×10^{-8} per base pair per generation (Schaffner *et al.* 2005) with a generation time of 25 years (Wich *et al.* 2009a).

We applied the following additional filtering to variant (SNP VCF files) and non-variant (gVCF files) sites for the calculation of the summary statistics: (i) genotypes with a sequence depth below 5-fold were excluded, and (ii) all sites had to be covered by at least two individuals per population of the respective species. For populations with sample size greater than four individuals, at least four called genotypes were required. Because of their low samples sizes (n = 2), we pooled the two Sabah populations North and South Kinabatangan. This can be justified as they were the last two clusters separating with increasing principal components in PCA, and data from the mitochondrial genome showed that they started to diverge only recently (Chapter 5).

Inference of demographic history

We inferred orangutan population size history with the pairwise sequentially Markovian coalescent (PSMC) model (Li & Durbin 2011), which uses single diploid genome sequences to reconstruct population size changes through time. The PSMC is implemented as a hidden Markov model in which the observation corresponds to the sequence of observed genotypes along the genome. The hidden state is the coalescent time of the two chromosomes at a given position, and transitions between hidden states represent ancestral recombination events. Thus, the PSMC model allows estimating historical changes in N_e based on the distribution of the time to the most recent common ancestor (TMRCA) for alleles within a diploid genome.

We applied the PSMC model to each sample. Input files for PSMC were created from the autosomal consensus FASTA sequences described above, using the utility 'fq2psmcfa' (provided with the PSMC package). We run PSMC with the following parameter settings, which were found to be suitable for great apes and applied to orangutans previously (Li & Durbin 2011; Prado-Martinez *et al.* 2013): 'psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o output.psmc input.psmcfa'. The parameter '-N' defines the number of iterations, '-t' the maximum TMRCA (in the 2N0 scale), '-r' the ratio of theta over rho, and '-p' describes the temporal binning parameters. In our case there were 64 atomic time intervals and 28 (=1+25+1+1) free interval parameters. We measured the variance of N_e estimates by bootstrapping. For each individual, we split its consensus sequence into 50-Mb segments using the 'splitfa' utility (PSMC package), and randomly sampled with replacement from these segments applying the '-b' option in PSMC for 100 rounds.

PSMC plots were drawn with a modified version of the 'psmc_plot.pl' script of the PSMC package. We scaled results to real time, assuming a generation time of 25 years (Wich *et al.* 2009a) and a mutation rate of 1.5×10^{-8} per site per generation (Schaffner *et al.* 2005) (Roach *et al.* 2010; The 1000 Genomes Project Consortium 2010). We generated different plots for high-coverage (≥ 20x), mid-coverage (11–18x), and low-coverage (5–6x) genomes as the trajectories of N_e should only be compared among genomes with similar read-depths. This is because the lower the coverage the higher the risk of missing a true heterozygous genotype, leading to reduced TMRCA in PSMC analyses (Li & Durbin 2011).

4.4 Results

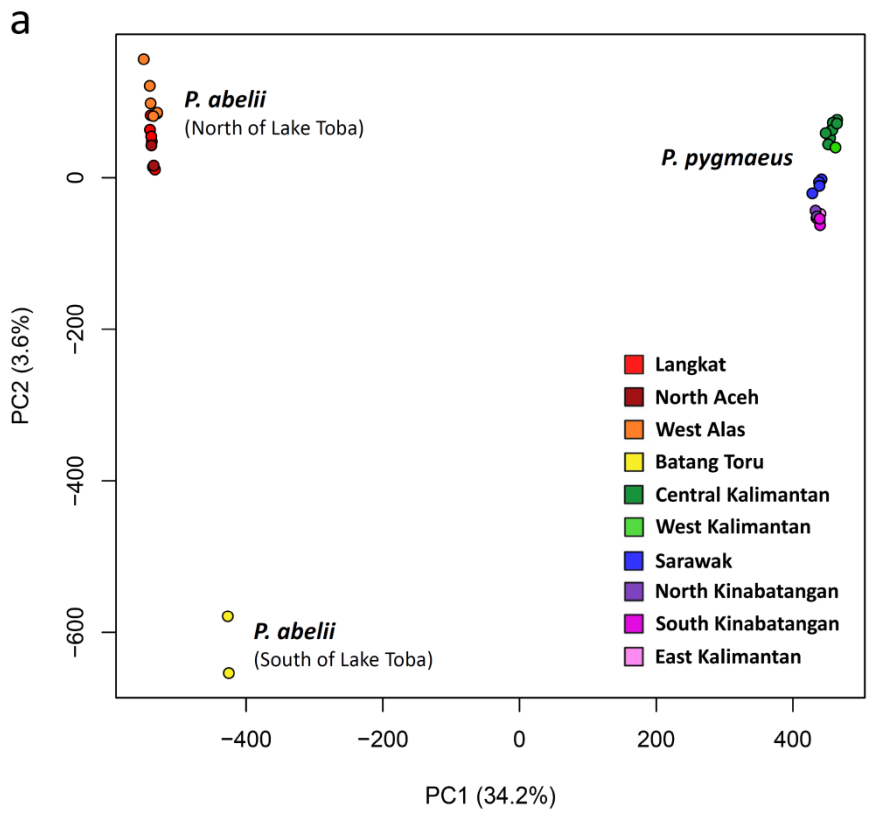
Whole-genome sequencing data

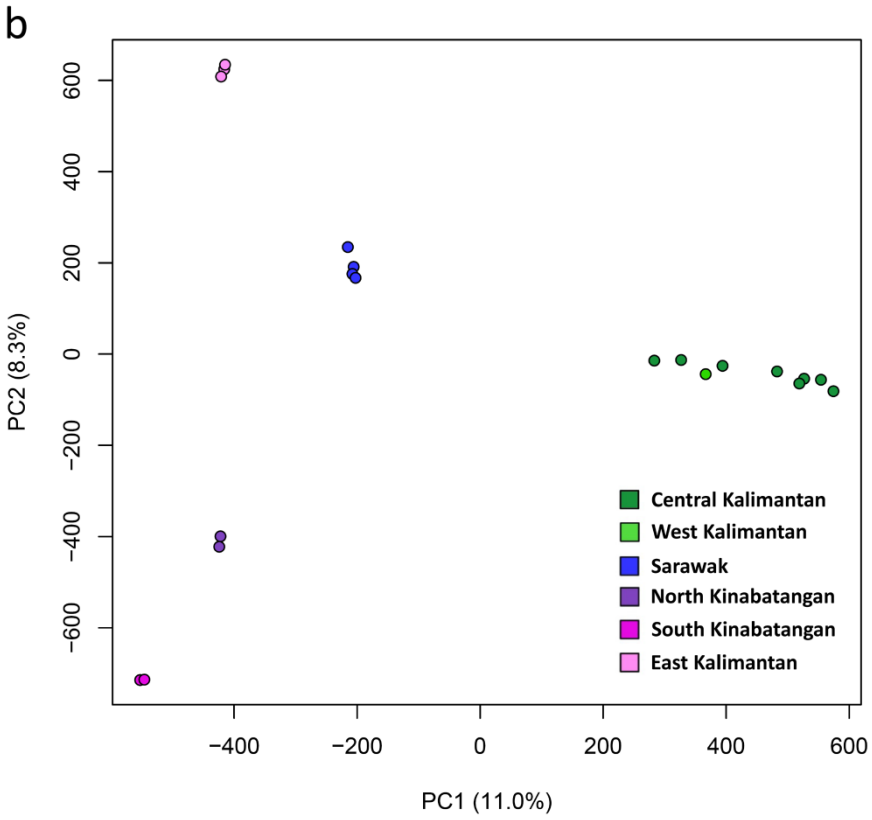
We complemented previous sequencing efforts by sequencing genomes of eleven wild-born orangutans with known provenance, generating on average ~$1.1x10^9$ raw Illumina reads per individual (Supporting Table S1). Basic sequencing and mapping statistics of all 36 study individuals are provided in the Supporting Tables S2 and S3. Mean effective sequence depth varied considerably among study individuals, ranging from 4.8x to 31.1x with an average depth of 18.4x over all individuals (Table 1, Supporting Table S1). For the previously sequenced genomes (Locke *et al.* 2011; Prado-Martinez *et al.* 2013), estimated sequence depths were 25–40% lower as the values reported in the two source studies. This difference is explained by the way sequence depth was calculated. Here, we estimated sequence depth on the filtered BAM files where duplicated reads, bad read mates, reads with mapping quality zero, and reads which mapped ambiguously had already been removed. Thus, our sequence coverage estimates correspond to the effective read-depths which are available for SNP discovery and genotyping. In total, we discovered 30,640,634 SNPs among all 36 individuals, which represents the most comprehensive catalogue of genetic diversity across the genus *Pongo* to date.

Population structure and genetic diversity

Principal component analysis revealed strong geographic structuring of the autosomal genetic diversity in the genus *Pongo* (Figure 2). As expected, in the analysis of all orangutans the first principal component (PC) separated the two orangutan species, explaining 34.2% of the total variance (Figure 2a). Surprisingly, however, the second PC isolated the two orangutans from Batang Toru from the other Sumatran populations north of Lake Toba (explaining 3.6% of the total variance).

Within species, the first two PCs partitioned autosomal genetic diversity into three distinct clusters in Sumatran orangutans (Figure 2b), and five main clusters in Bornean orangutans (Figure 2c). Notably, the five captive-born individuals clustered autosomally with their natal population, thus do not seem to be within-species hybrids. The identified population structure is in agreement with orangutan mtDNA phylogeography (Nater *et al.* 2011) and matches the previously described genetic clusters from classical microsatellite markers (Chapter 3; Arora *et al.* 2010; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015). We found that North Aceh and Langkat do not form autosomally distinct clusters, despite their deep mitochondrial divergence of ~0.85 Ma (Nater *et al.* 2011). Thus, they seem to represent two sub-populations (both located on the same side of the Alas River), which remained connected by considerable male-driven gene flow after their initial separation.


a
P. abelii
(North of Lake Toba)
P. pygmaeus
P. abelii
(South of Lake Toba)
Langkat
North Aceh
West Alas
Batang Toru
Central Kalimantan
West Kalimantan
Sarawak
North Kinabatangan
South Kinabatangan
East Kalimantan
PC2 (3.6%)
PC1 (34.2%)
0
−200
−400
−600
−400
−200
0
200
400


b
Central Kalimantan
West Kalimantan
Sarawak
North Kinabatangan
South Kinabatangan
East Kalimantan
PC2 (8.3%)
PC1 (11.0%)
600
400
200
0
−200
−400
−600
−400
−200
0
200
400
600

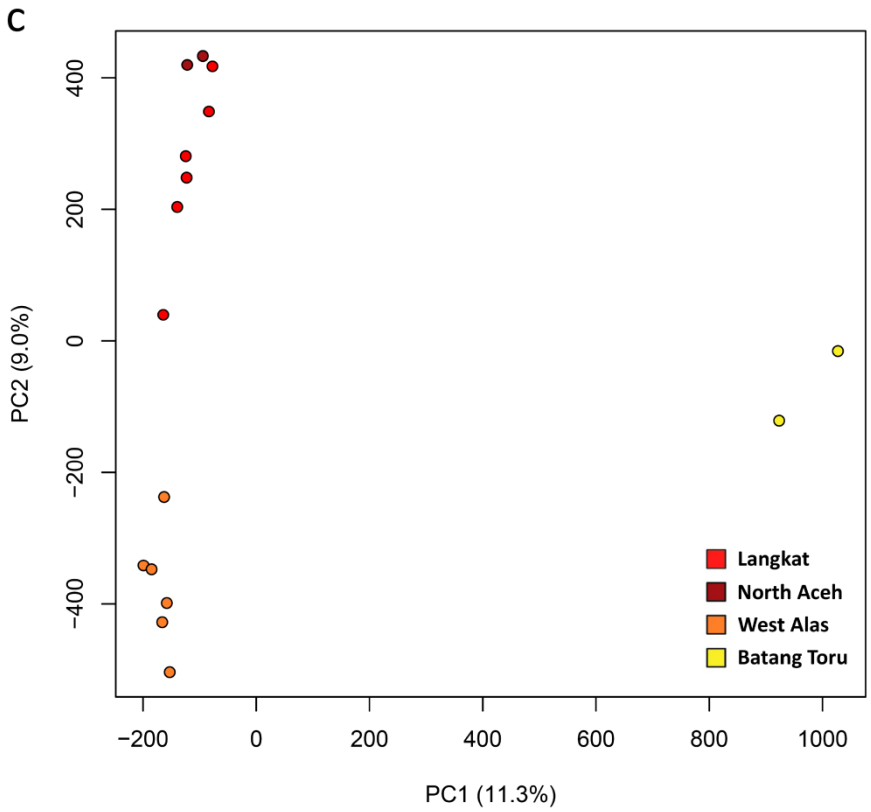


Figure 2. Principal component analysis of the genetic diversity present in autosomal genomes of (a) all orangutans, (b) Bornean orangutans, and (c) Sumatran orangutans. Each dot represents an individual. Color codes match those of Figure 1. The percentage of the total variance explained by a principal component (PC) is given in parenthesis.

Genome-wide diversity was on average ~26% higher in Sumatran orangutans than in Bornean orangutans (Table 2). Effective population sizes were similar for all populations within species, ranging from N_e [θ_π] = 39,345–41,511 for Sumatran populations and N_e [θ_π] = 28,778–30,211 for Bornean populations. Generally, averaged individual heterozygosity values and pairwise nucleotide diversities among individuals within populations were in congruence.

Table 2. Genetic diversity by population. Because of low sample sizes and only recent divergence (Chapter 5), we pooled Individuals from South and North Kinabatangan.

	$N_{Samples}$	Average depth[a]	Heterozygosity (bp^{-1})[b]	θ_π[c]	θ_W[d]	N_e [θ_π][e]	N_e [θ_W][e]
P. abelii	16	16.1	0.0024	0.0026	0.0025	43,366	41,204
Langkat/North Aceh	8	19.0	0.0025	0.0025	0.0024	41,511	39,944
West Alas	6	14.0	0.0023	0.0024	0.0024	40,144	39,182
Batang Toru	2	11.4	0.0022	0.0024	0.0023	39,345	38,986
P. pygmaeus	20	20.2	0.0017	0.0019	0.0017	31,869	27,501
Central/West Kalimantan	9	18.3	0.0018	0.0018	0.0016	29,899	26,179
East Kalimantan	3	22.0	0.0018	0.0017	0.0017	28,778	27,641
Sarawak	4	17.7	0.0018	0.0018	0.0017	30,211	28,811
South/North Kinabatangan	4	25.7	0.0017	0.0018	0.0017	30,190	28,269

[a]average effective sequencing depth (estimated from the filtered BAM files; see Materials and Methods)
[b]heterozygotes per base pair (bp)
[c]Theta from π
[d]Watterson estimator θ_W
[e]long-term effective population size calculated from $\theta_{\pi/W}$

Demographic history

We inferred historical changes of autosomal N_e using the PSMC model (Figure 3, Supporting Figures S4–6). Trajectories of the PSMC suggest that Bornean and Sumatran orangutans diverged ~0.9–1.1 Ma (scaling 0.6 x10^{-9} per base pair per year; Figure 3). Subsequently, the two species experienced very different demographic histories. Bornean orangutans underwent an initial population decline followed by short recovery. Around 300 ka, N_e began to decline continuously, resulting in very low N_e in the more recent past. In contrast, N_e of Sumatran orangutans increased considerably after species separation, which could represent actual population growth, a signal of increased population sub-structuring, or most likely a combination of both (see Discussion). More recently (50–100 ka) autosomal N_e of Sumatran orangutans dropped sharply to just a few thousand individuals, coinciding with the Toba supereruption ~73 ka (Chesner *et al.* 1991).

Within species, the trajectories of N_e were highly similar for individuals across populations as shown by the overlap of the variance of their PSMC estimates (Supporting Figure S4). Notably, the Batang Toru population is missing in Figure 3 because effective sequence read-depth for both individuals was below the required 20-fold threshold. However, their N_e trajectories also fell together with those of other Sumatran orangutans having similar read-depths (Supporting Figures S5 and S6).

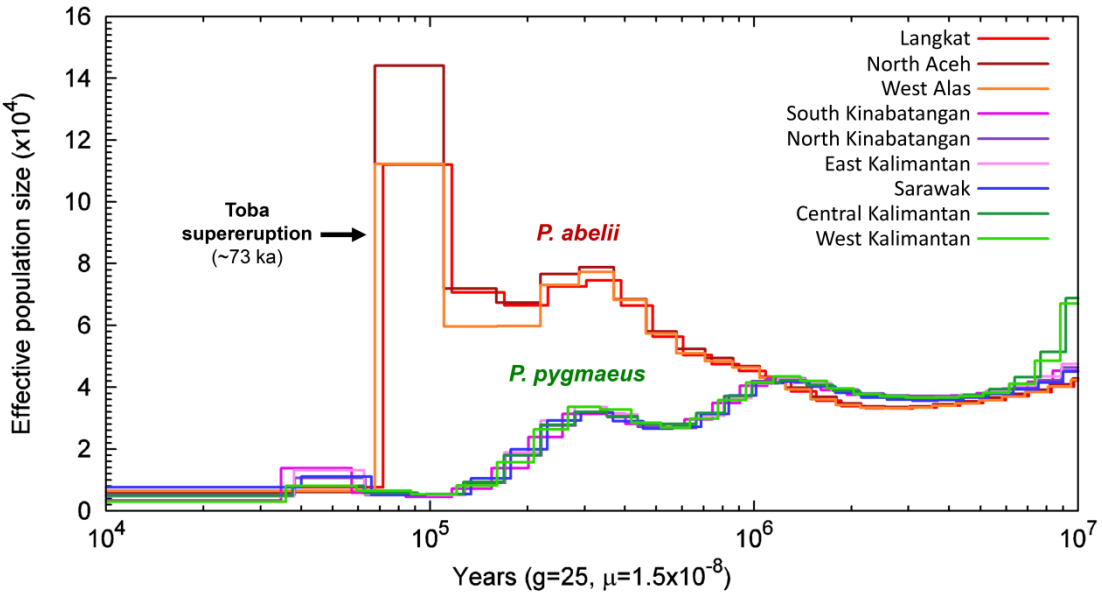


Figure 3. Demographic history of the genus *Pongo*. Autosomal N_e history was inferred by pairwise sequentially Markovian coalescent (PSMC) analysis. For each orangutan population, one high-coverage (≥ 20x) genome is plotted (except Batang Toru, see main text). Color codes match those of Figure 1. The x-axis gives time scaled in years, assuming a generation time of 25 years and a mutation rate of 1.5x10^{-8} per site per generation. The y-axis shows historical Ne. Note that the PSMC model cannot detect N_e changes more recent than 20,000 years ago due to few recent coalescent events. Also, sudden changes of N_e tend to be smoothed out (Li & Durbin 2011).

4.5 Discussion

We investigated the autosomal demographic history and extant population structure of the genus *Pongo* using the most extensive genomic sampling to date, encompassing genomes of 36 unrelated orangutans and covering the entire current geographic range of the genus. This was achieved by complementing previous sequencing efforts (Locke *et al.* 2011; Prado-Martinez *et al.* 2013) by sequencing genomes of eleven wild-born orangutans from missing or underrepresented populations to an average 23-fold coverage per individual.

Species divergence

The PSMC analysis suggests that autosomal gene pools of Sumatran and Bornean orangutans diverged ~0.9–1.1 Ma (scaling 0.6 $x10^{-9}$ $bp^{-1}y^{-1}$). Yet, data from mtDNA indicate that initial separation already might have taken place as early as the beginning of the Pleistocene ~2.0–2.5 Ma (Chapter 5; Nater *et al.* 2011). This discrepancy illustrates that orangutan speciation has been a gradual process over several hundred thousand years during which autosomal gene pools of Bornean and Sumatran orangutans remained connected via exclusively male-mediated gene flow after their initial separation.

The inferred autosomal divergence time may not correspond to the final cessation of gene flow between the two islands, although rates were at least substantially reduced. However, considering the strikingly different PSMC profiles of Bornean and Sumatran orangutans, it seems unlikely that they were connected by gene flow during the last two glacial periods (i.e. the last 200,000 years) as proposed previously (Muir *et al.* 2000; Verschoor *et al.* 2004; Steiper 2006; Becquet & Przeworski 2007; Nater *et al.* 2011; Nater *et al.* 2015). For instance, large river systems dissecting the exposed Sunda Shelf at low sea levels during glacial periods (Rijksen & Meijaard 1999; Harrison *et al.* 2006) and a savanna corridor (Gathorne-Hardy *et al.* 2002; Bird *et al.* 2005; Slik *et al.* 2011) may have imposed significant dispersal barriers.

Demographic history Bornean orangutans

After their divergence, Bornean and Sumatran orangutans experienced very different autosomal demographic histories. The initial population decline with subsequent expansion and the strong continuous decline in the late Pleistocene observed in Bornean orangutans are most likely linked to the Quaternary climatic oscillations. Glacial periods were considerably drier and more seasonal than inter-glacials, leading to repeated rainforest contractions and expansions (Flenley 1998; Morley 2000; Bird *et al.* 2005). It is therefore conceivable that Bornean orangutans experienced a series of population bottlenecks with subsequent expansions during the Pleistocene which we cannot detect because sudden changes of N_e tend to be smoothed out in the PSMC model (Li & Durbin 2011). Such a scenario is supported by the recent coalescence of Bornean mtDNA lineages, providing evidence for a population bottleneck during the penultimate glacial period (130–190 ka) (Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2015).

Demographic history Sumatran orangutans and the Toba supereruption

In contrast to the initial population decline of Bornean orangutans after the autosomal species divergence, N_e of Sumatran orangutans increased substantially. This increase could be linked to range expansion during the Middle Pleistocene. However, it may also partly represent an artefact from increased population structuring, for instance as a consequence of reduced gene flow among populations after volcanic eruptions or during glacial periods when rainforests were contracted to refugia (Gathorne-Hardy *et al.* 2002). Because rain-fall rates during glacials were considerably higher for most of Sumatra comparted to Borneo, multiple rainforest refugia likely existed along the Barisan Mountain range and in northern Sumatra (Gathorne-Hardy *et al.* 2002). The isolation among the different refugia may have led to increased population structuring of gene pools in Sumatran orangutans. Although the trajectories of N_e are inferred from individual genomes in the PSMC model, the historical N_e estimates reflect the size of the entire meta-population, as lineages might migrate among subpopulations before they coalesce. This also explains the congruence of N_e trajectories observed for different populations within species in PSMC. According to coalescence theory, N_e of the meta-population is larger than the sum of N_e of all sub-populations since coalescence occurs less frequently in structured populations (Hudson 1990). Thus, increasing population structuring may lead to an expansion signal in PSMC analysis.

The most striking observation of the PSMC analysis was the enormous abrupt drop of autosomal N_e of Sumatran orangutans around the time of the Toba supereruption ~73 ka (Chesner *et al.* 1991). The steep increase in N_e before this event might reflect an artefact of the PSMC model in consequence of the sudden drastic change in N_e. Because N_e of Bornean orangutans was already very small at that time, we lack resolution to make inferences on consequences of the Toba supereruption on Bornean orangutans. Linking changes of N_e in PSMC with specific environmental events is difficult because of uncertainties associated with the applied substitute rate to scale results to real time. Thus, more detailed studies are required to confirm our findings.

To our knowledge, the observed drastic population decline of Sumatran orangutans is the first direct evidence of a strong regional impact of the Toba supereruption on a large mammal, and implies that the consequences of the supereruption on fauna and flora may have been more severe than argued previously (Schulz *et al.* 2002; Gathorne-Hardy & Harcourt-Smith 2003; Petraglia *et al.* 2007; Haslam & Petraglia 2010). Despite this drastic impact, however, it is unlikely that habitat was completely destroyed over large areas as proposed earlier by others (Rampino & Ambrose 2000; Williams *et al.* 2009). Under such a scenario, we would not observe a persistence of multiple old maternal lineages in Sumatran orangutans in close proximity to the Toba caldera as well as the paraphyly of Sumatran orangutan mtDNA lineages (Chapter 5; Nater *et al.* 2011).

Overall, our PSMC results provide strong evidence against the demographic model of one continuously expanding panmictic Sumatran population with large current N_e inferred by

Locke *et al.* (2011). As already reported by Nater *et al.* (2015), this study shows that the misleading signals of Locke *et al.* resulted from ignoring the deep structuring of Sumatran populations and their strong recent population decline. Sampling bias and neglect of population structure provide misleading results of genetic diversity and temporal changes of N_e (Stadler *et al.* 2009; Chikhi *et al.* 2010; Peter *et al.* 2010). A representative sampling covering the entire range of a species is therefore critical for accurate reconstruction of demographic history (Stadler *et al.* 2009), in particular if populations are as deeply structured as in orangutans.

Special status of the Batang Toru population and conservation implications

Using samples from the entire extant range of orangutans, we identified three autosomally distinct orangutan populations for *P. abelii* in Sumatra and five for *P. pygmaeus* in Borneo. Our most striking finding was the highly distinct separation of the Sumatran populations north of Lake Toba and Batang Toru, the only remnant population south of the extensive caldera. This corresponds well with data from classical autosomal microsatellite makers for which genetic differentiation among Sumatran orangutans was also highest across Lake Toba (Nater *et al.* 2013). A particularly deep divergence of orangutans north and south of Lake Toba has also been found for mtDNA genes. Nater *et al.* (2011) observed the oldest split (around 3.5 Ma) in the mtDNA phylogeny of the genus *Pongo* between the lineage leading to Batang Toru and all Bornean orangutans, and that of Sumatran populations north of Lake Toba, rather than between Bornean and Sumatran orangutans. The extensive volcanic activity of Mount Toba (Chesner *et al.* 1991) seems to have led to a long-lasting separation of orangutan gene pools south and north of the Toba caldera. This finding is well in line with the fact that Lake Toba also represents a significant zoogeographic boundary for many other species, including birds (Whitten 2000), Lar and Agile gibbons (Whittaker *et al.* 2007; Thinh *et al.* 2010), and the Thomas's langurs (Aimi & Bakar 1996).

Our finding that the Batang Toru population is autosomally highly distinct from all other Sumatran populations and separates as a third cluster after the two recognized species, has important ramifications for conservation and taxonomy. Due to their dependency on intact rainforest and their exceptionally slow life history, orangutans are severely affected by ongoing habitat destruction and fragmentation as well as illegal hunting. Sumatran orangutans are listed as critically endangered and Bornean orangutans as endangered (IUCN 2014), with only an estimated 6,600 Sumatran and 54,000 Bornean orangutans left in the wild (Wich *et al.* 2008). The Batang Toru population is particularly threatened because most of the forest in this area is not under protection (Wich *et al.* 2011a; Wich *et al.* 2014). Given the genetic uniqueness of the orangutans in this area (Nater *et al.* 2011; Nater *et al.* 2015), we strongly propose to protect them as a separate evolutionary significant unit (ESU) and suggest a taxonomic revision of *P. abelii*. Considering their already extremely low current census size of only 400-600 individuals (Wich *et al.* 2008; Marshall *et al.* 2009), urgent actions need to be taken to preserve this indispensable reservoir of genetic diversity of orangutans.

In conclusion, our study showed that orangutans experienced a complex demographic history and exhibit deep population structure in the autosomal genome, which both need to be taken into account when studying the adaptive evolution of *Pongo* and great apes in general. The whole-genome data generated in this study will serve as a unique resource for further research, such as tackling the genetic basis of the remarkable geographic variation of phenotypic traits in these fascinating Asian great apes.

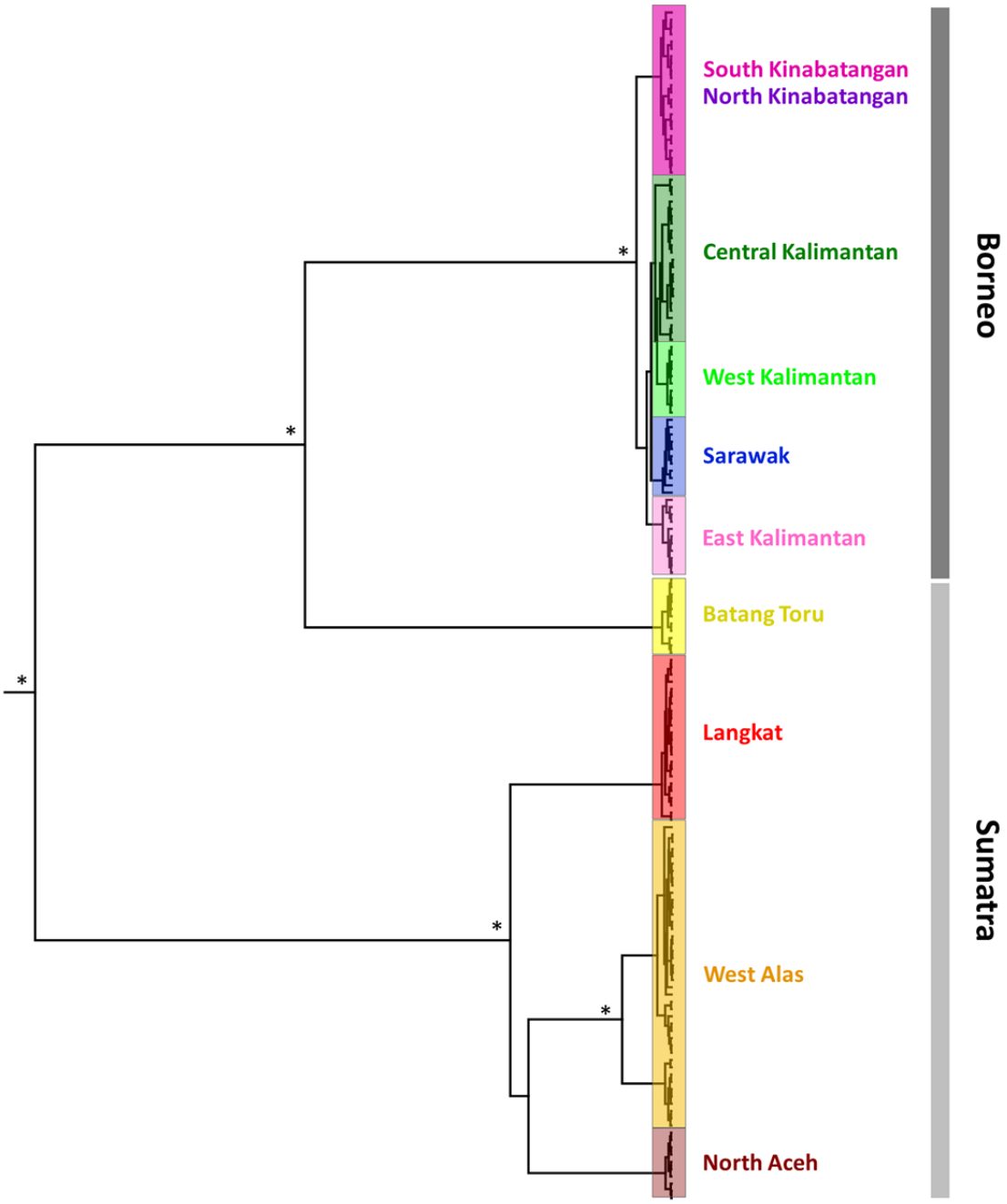
Acknowledgments

We thank the Centre Nacional d'Anàlisi Genòmica and the Functional Genomics Centre Zurich for performing whole-genome sequencing. We are grateful to Joko Pamungkas, Dyah Perwitasari-Farajallah, Muhammad Agil as well as the staff at the Sumatran Orangutan Conservation Programme, Sepilok Orangutan Rehabilitation Centre, BOS Wanariset Orangutan Reintroduction Project, and Semongok Wildlife Rehabilitation Centre who helped collecting and exporting samples. Furthermore, we thank the following institutions for supporting our research: Sabah Wildlife Department (SWD), Indonesian State Ministry for Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), Leuser International Foundation (LIF), Taman National Gunung Leuser (TNGL), and Borneo Orangutan Survival Foundation (BOSF). This study was financially supported by UZH University Research Priority Program, Leakey Foundation (to MPG), ERC Starting Grant (grant no. 260372 to TMB), Swiss National Science Foundation (grant no. 3100A-116848 to MK and CPvS), Forschungskredit University of Zurich (to MPG), Julius–Klaus Foundation (to MK), A.H. Schultz Foundation (to MK and MPG), and the Anthropological Institute & Museum at the University of Zurich.

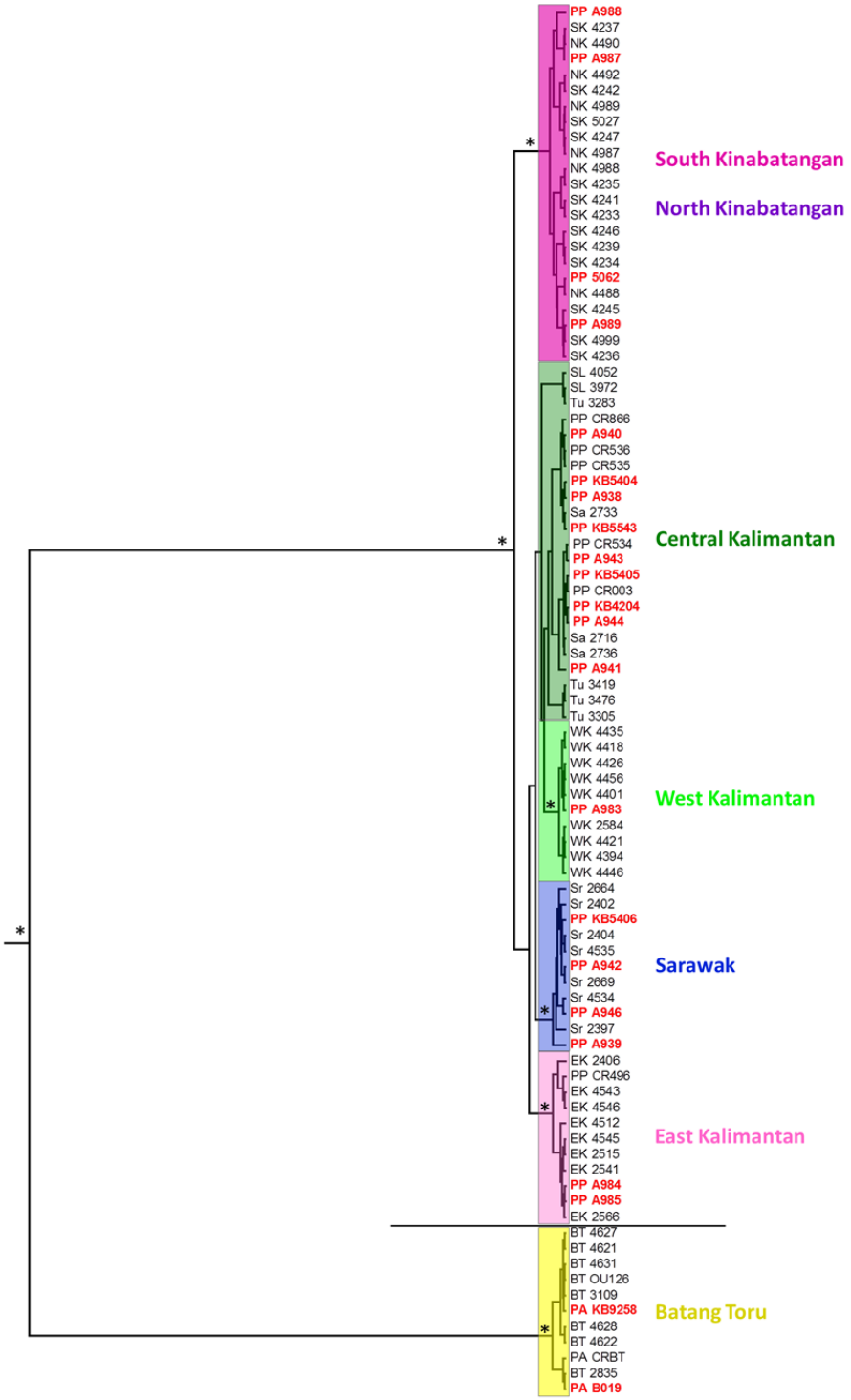
Author Contributions

MPG, AN, TMB, and MK conceived the study. MPG and MK coordinated the study. BG, MPG, MK, EV, KS, IS, AN, LNA, and CPvS provided genetic samples. MPG performed population genetic assessment of study individuals. IG and MG carried out sequencing. TMB and JPM contributed additional sequencing data. AN and MPG performed short read mapping, SNP and genotype calling. MPG conducted statistical analyses. AN provided novel bioinformatic tools and conducted statistical analyses. MPG wrote the manuscript. AN and MK edited the manuscript.

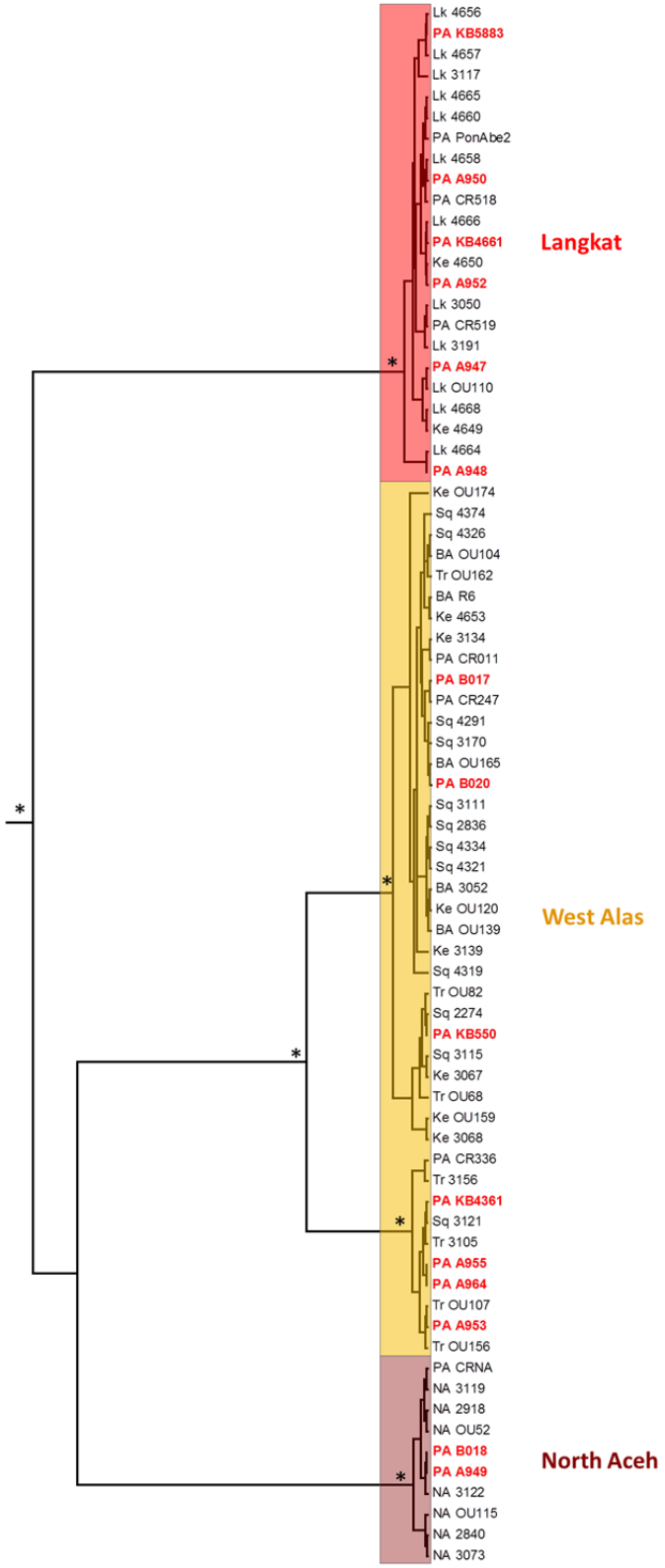
4.6 Supporting Information



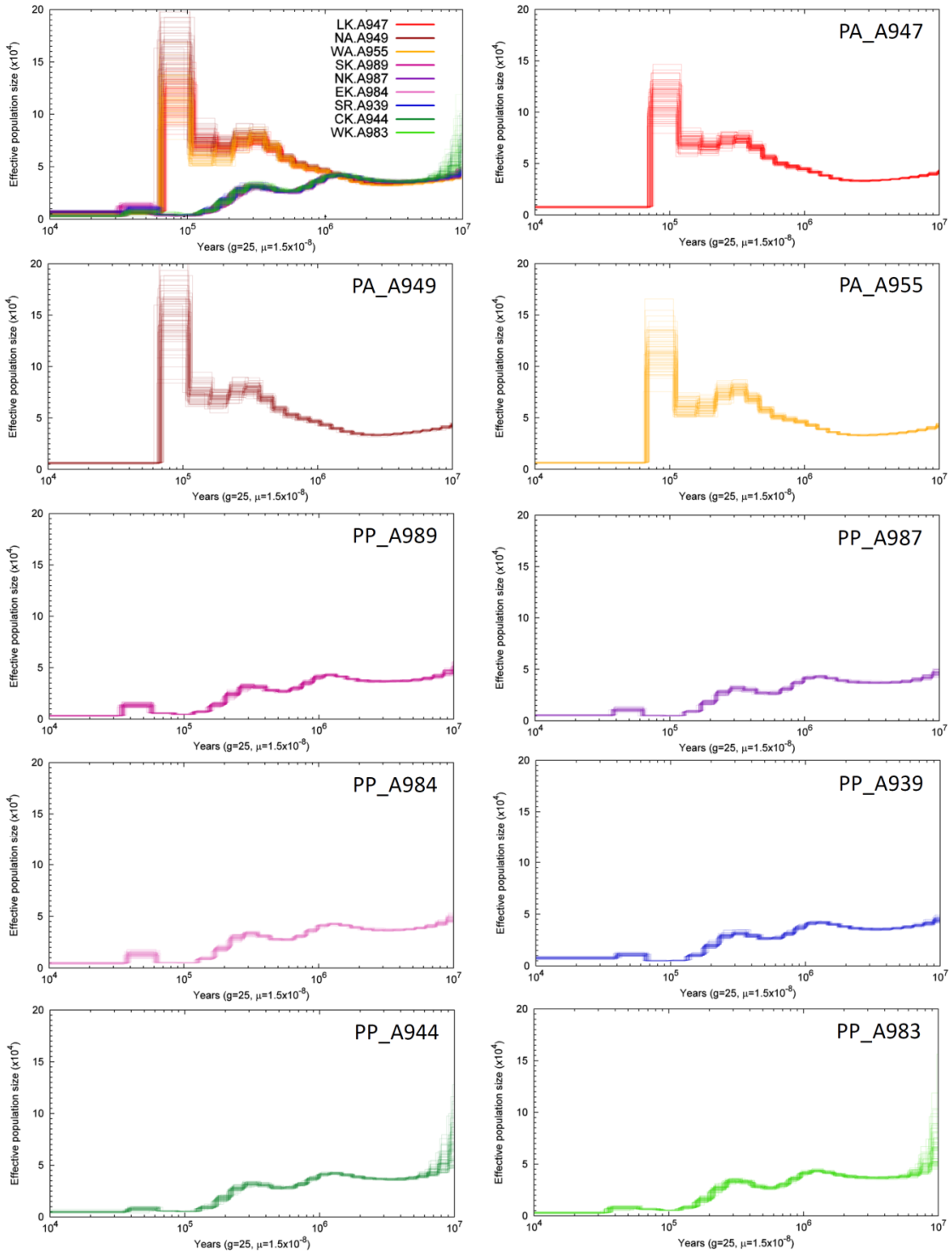
Supporting Figure S1. Bayesian phylogenetic mtDNA tree of the genus *Pongo*. The tree was derived from three concatenated mtDNA genes and includes 127 non-invasively sampled wild orangutans covering the entire range of extant orangutans (Nater *et al.* 2011) as well as the 36 study individuals. Color codes of orangutan populations match those of Figure 1. The posterior probability of each clade was >97% (*). Nodes within the color shaded areas are not annotated due to space constrains. The tree is rooted with a human and a central chimpanzee sequence (not shown). Sub-trees for close-up view are provided in Supporting Figures S2 and S3.



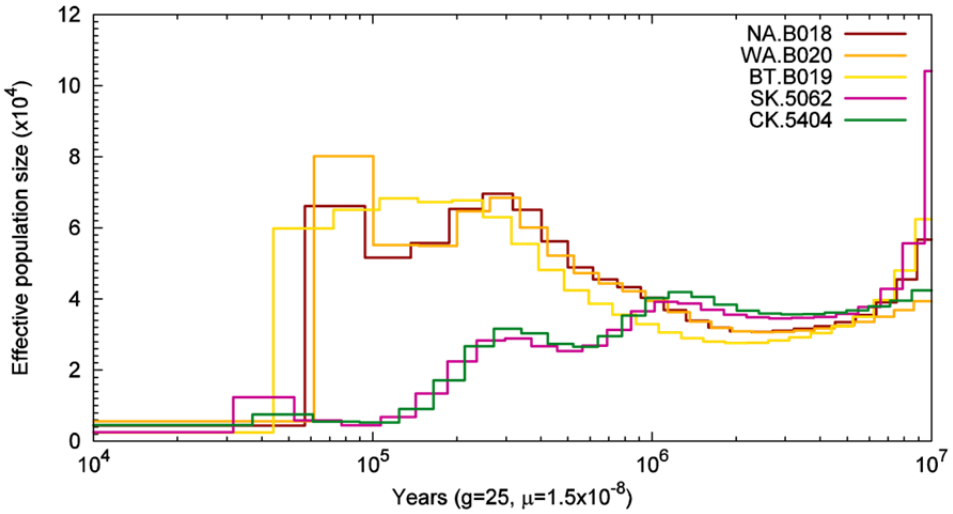
Supporting Figure S2. Sub-tree of Bornean orangutans and the Sumatran Batang Toru population. The sub-tree was extracted from the maximum clade credibility tree in Supporting Figure S1. Study individuals are highlighted in red. Color codes of orangutan populations match those of Figure 1. The posterior probability (PP) of each relevant clade was >97% (*), except for Central Kalimantan (PP 34%). South and North Kinabatangan did not form reciprocally monophyletic groups for the three mtDNA genes underlying this tree. However, none of the study individuals with unknown provenance felt within this cluster.



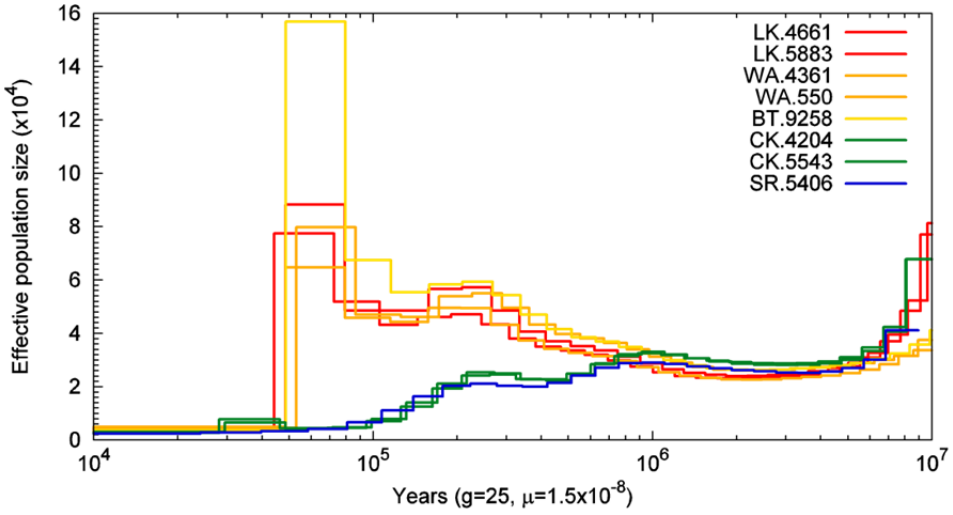
Supporting Figure S3. Sub-tree of Sumatran orangutans north of Lake Toba. The sub-tree was extracted from the maximum clade credibility tree in Supporting Figure S1. Study individuals are highlighted in red. Color codes of orangutan populations match those of Figure 1. The posterior probability of each clade was >97% (*).



Supporting Figure S4. PSMC bootstrapping plots for individuals in Figure 3. The x-axis gives time scaled in years, assuming a generation time of 25 years and a mutation rate of 1.5x10-^{8} per site per generation. The y-axis shows historical Ne. The fluctuation of the 100 bootstrap replicates indicates the variance. The plot on the top left shows the overlay of all nine individuals. Color codes match those of Figure 1.



Supporting Figure S5. PSMC analysis of mid-coverage (11–18x) orangutan genomes. The x-axis gives time scaled in years, assuming a generation time of 25 years and a mutation rate of 1.5×10^{-8} per site per generation. The y-axis shows historical Ne. Color codes match those of Figure 1. Details on individuals can be found in Supporting Table S1.



Supporting Figure S6. PSMC analysis of low-coverage (5–6x) orangutan genomes. The x-axis gives time scaled in years, assuming a generation time of 25 years and a mutation rate of 1.5×10^{-8} per site per generation. The y-axis shows historical Ne. Color codes match those of Figure 1. Details on individuals can be found in Supporting Table S1.

Supporting Table S1. Details on study individuals.

Species	Population	Individual ID	Individual name	Sex	Mean depth[a]	Source	Comments and origin details
P. abelii	Langkat	PA_KB4661	Bubbles	M	4.76	Locke *et al.* 2011	Wild-born
P. abelii	Langkat	PA_KB5883	Sibu	M	4.99	Locke *et al.* 2011	Wild-born
P. abelii	Langkat	PA_A947	Elsi	F	27.39	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A948	Kiki	F	23.71	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A950	Babu	F	26.28	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A952	Buschi	M	21.03	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	North Aceh	PA_A949	Dunja	F	27.39	Prado-Martinez *et al.* 2013	1. Gen. by 456 and 457 both wild-born Sumatra
P. abelii	North Aceh	PA_B018	Jeff	M	16.31	this study	Wild-born; Desa Seuneubok Bayu, Kec. Indra Makmu; ID AIM: 5295
P. abelii	West Alas	PA_KB4361	Likoe	F	5.66	Locke *et al.* 2011	Wild-born
P. abelii	West Alas	PA_SB550	Doris	F	4.86	Locke *et al.* 2011	Wild-born
P. abelii	West Alas	PA_B017	Miky	f	13.74	this study	Wild-born; Aluebillie, Aceh Nagan Raya, Aceh province; ID AIM: 5252
P. abelii	West Alas	PA_A953	Vicky	F	17.78	unpubl. Prado-Martinez *et al.* 2013	Wild-born
P. abelii	West Alas	PA_A955	Suma	F	25.27	unpubl. Prado-Martinez *et al.* 2013	Wild-born
P. abelii	West Alas	PA_B020	Maini	F	16.3	this study	Wild-born; Aceh Sealatan near Suaq Balimbing; ID AIM: 3111
P. abelii	Batang Toru	PA_KB9258	Baldy	F	5.79	Locke *et al.* 2011	Wild-born
P. abelii	Batang Toru	PA_B019	Afa	M	16.92	this study	Wild-born; ID AIM: 2835
P. pygmaeus	Central Kalimantan	PP_KB4204	Dolly	M	5.61	Locke *et al.* 2011	Wild-born
P. pygmaeus	Central Kalimantan	PP_KB5404	Billy	F	12.24	Locke *et al.* 2011	Wild-born
P. pygmaeus	Central Kalimantan	PP_KB5405	Dennis	M	5.61	Locke *et al.* 2011	Wild-born
P. pygmaeus	Central Kalimantan	PP_A940	Temmy	F	21.8	Prado-Martinez *et al.* 2013	1. Gen. by 793 and 794 both wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_A941	Sari	F	23.17	Prado-Martinez *et al.* 2013	1. Gen. by 202 and 322 both wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_A943	Tilda	F	24.17	Prado-Martinez *et al.* 2013	Wild-born
P. pygmaeus	Central Kalimantan	PP_A944	Napoleon	M	23.32	Prado-Martinez *et al.* 2013	Wild-born
P. pygmaeus	Central Kalimantan	PP_A938	Lotti	F	18.62	unpubl. Prado-Martinez *et al.* 2013	1. Gen. by 358 and 422 both wild-born Borneo
P. pygmaeus	West Kalimantan	PP_A983	Claus	M	29.71	this study	Wild-born; Pontianak; ID AIM: 4560
P. pygmaeus	East Kalimantan	PP_KB5543	Louis	M	6.03	Locke *et al.* 2011	Wild-born
P. pygmaeus	East Kalimantan	PP_A984	Barong	F	29.89	this study	Wild-born; Taman Nasional Kutail; ID AIM: 4552
P. pygmaeus	East Kalimantan	PP_A985	Panjul	M	30.13	this study	Wild-born; Taman Nasional Kutail; ID AIM: 4592
P. pygmaeus	North Kinabatangan	PP_A987	Tara	F	30.65	this study	Wild-born; Bukit Garam, Kinabatangan area; ID AIM: 5044
P. pygmaeus	North Kinabatangan	PP_A988	Kala	M	31.06	this study	Wild-born; Kg. Tikolod, Tambunan; ID AIM: 5053
P. pygmaeus	South Kinabatangan	PP_5062	Ampal	M	13.81	this study	Wild-born; Lahad Datu, Kinabatangan area; ID AIM: 5062
P. pygmaeus	South Kinabatangan	PP_A989	Micelle	F	27.30	this study	Wild-born; Lahad Datu, Kinabatangan area; ID AIM: 5057
P. pygmaeus	Sarawak	PP_KB5406	Dinah	F	4.90	Locke *et al.* 2011	Wild-born
P. pygmaeus	Sarawak	PP_A939	Nonja	F	20.48	Prado-Martinez *et al.* 2013	1. Gen. by 1052 and 1012 both from Sarawak
P. pygmaeus	Sarawak	PP_A942	Gusti	F	23.12	unpubl. Prado-Martinez *et al.* 2013	1. Gen. by 1435 and 1392 both wild-born Borneo
P. pygmaeus	Sarawak	PP_A946	Kajan	M	22.39	unpubl. Prado-Martinez *et al.* 2013	Wild-born

[a]mean effective whole-genome sequencing coverage (estimated from the filtered BAM files, see Materials and Methods)

Supporting Table S2. Basic sequencing and mapping statistics of orangutan whole-genome sequencing data.

Species	Individual ID	Source	Total no. of reads	No. of reads filtered out	% reads filtered out	No. of bad mate reads[a]	% bad mate reads
P. abelii	PA_A947	Prado-Martinez *et al.* 2013	1,199,070,495	217,651,201	18.15%	31,965,745	2.67%
P. abelii	PA_A948	Prado-Martinez *et al.* 2013	1,026,568,611	212,620,172	20.71%	23,791,092	2.32%
P. abelii	PA_A949	Prado-Martinez *et al.* 2013	1,238,435,940	295,572,494	23.87%	28,597,946	2.31%
P. abelii	PA_A950	Prado-Martinez *et al.* 2013	1,221,075,045	277,425,024	22.72%	26,033,305	2.13%
P. abelii	PA_A952	Prado-Martinez *et al.* 2013	1,061,059,740	333,654,395	31.45%	26,183,487	2.47%
P. abelii	PA_A953	unpubl. Prado-Martinez *et al.* 2013	863,795,942	240,805,194	27.88%	16,835,303	1.95%
P. abelii	PA_A955	unpubl. Prado-Martinez *et al.* 2013	1,151,082,160	258,616,853	22.47%	25,494,067	2.21%
P. abelii	PA_B017	this study	1,114,451,019	576,916,768	51.77%	320,927,625	28.80%
P. abelii	PA_B018	this study	1,213,126,904	606,523,688	50.00%	380,058,442	31.33%
P. abelii	PA_B019	this study	1,065,556,174	447,571,547	42.00%	218,524,547	20.51%
P. abelii	PA_B020	this study	1,063,963,834	467,186,672	43.91%	268,296,390	25.22%
P. abelii	PA_KB4361	Locke *et al.* 2011	502,515,251	102,527,136	20.40%	5,668,056	1.13%
P. abelii	PA_KB4661	Locke *et al.* 2011	395,184,293	76,284,313	19.30%	4,802,843	1.22%
P. abelii	PA_KB5883	Locke *et al.* 2011	470,563,961	115,172,006	24.48%	7,246,997	1.54%
P. abelii	PA_KB9258	Locke *et al.* 2011	546,292,437	118,733,071	21.73%	6,956,518	1.27%
P. abelii	PA_SB550	Locke *et al.* 2011	420,906,050	87,518,248	20.79%	6,816,380	1.62%
P. pygmaeus	PP_5062	this study	520,463,882	71,616,442	13.76%	12,238,321	2.35%
P. pygmaeus	PP_A938	unpubl. Prado-Martinez *et al.* 2013	878,679,380	219,613,306	24.99%	18,771,601	2.14%
P. pygmaeus	PP_A939	Prado-Martinez *et al.* 2013	982,875,157	258,243,405	26.27%	22,266,716	2.27%
P. pygmaeus	PP_A940	Prado-Martinez *et al.* 2013	879,365,509	111,712,294	12.70%	23,951,485	2.72%
P. pygmaeus	PP_A941	Prado-Martinez *et al.* 2013	974,172,871	162,961,808	16.73%	22,324,151	2.29%
P. pygmaeus	PP_A942	unpubl. Prado-Martinez *et al.* 2013	1,119,665,510	294,172,924	26.27%	27,378,538	2.45%
P. pygmaeus	PP_A943	Prado-Martinez *et al.* 2013	1,137,225,178	276,513,275	24.31%	28,140,416	2.47%
P. pygmaeus	PP_A944	Prado-Martinez *et al.* 2013	1,110,367,688	280,618,436	25.27%	30,240,109	2.72%
P. pygmaeus	PP_A946	unpubl. Prado-Martinez *et al.* 2013	944,435,510	165,822,299	17.56%	19,510,440	2.07%
P. pygmaeus	PP_A983	this study	1,150,227,749	171,282,032	14.89%	27,964,164	2.43%
P. pygmaeus	PP_A984	this study	1,166,011,497	181,228,288	15.54%	32,704,080	2.80%
P. pygmaeus	PP_A985	this study	1,188,314,591	190,300,804	16.01%	38,933,010	3.28%
P. pygmaeus	PP_A987	this study	1,182,067,514	169,028,331	14.30%	32,242,622	2.73%
P. pygmaeus	PP_A988	this study	1,184,387,913	159,637,530	13.48%	28,471,644	2.40%
P. pygmaeus	PP_A989	this study	1,182,468,671	254,009,220	21.48%	111,433,632	9.42%
P. pygmaeus	PP_KB4204	Locke *et al.* 2011	488,513,841	91,445,743	18.72%	5,431,871	1.11%
P. pygmaeus	PP_KB5404	Locke *et al.* 2011	1,223,090,264	279,931,929	22.89%	26,711,713	2.18%
P. pygmaeus	PP_KB5405	Locke *et al.* 2011	450,850,553	102,845,293	22.81%	3,703,627	0.82%
P. pygmaeus	PP_KB5406	Locke *et al.* 2011	427,501,183	79,470,592	18.59%	5,199,834	1.22%
P. pygmaeus	PP_KB5543	Locke *et al.* 2011	531,449,862	133,019,779	25.03%	7,881,803	1.48%

Supporting Table S2 (Continued)

Species	Individual ID	No. of duplicate reads	% duplicate reads	No. of MappingQualityZero reads	% MappingQualityZero reads	No. of NotPrimaryAlignment reads[b]	% NotPrimaryAlignment reads
P. abelii	PA_A947	101,285,592	8.45%	83,954,851	7.00%	445,013	0.04%
P. abelii	PA_A948	110,628,113	10.78%	77,752,127	7.57%	448,840	0.04%
P. abelii	PA_A949	185,752,688	15.00%	80,721,416	6.52%	500,444	0.04%
P. abelii	PA_A950	121,393,826	9.94%	129,564,956	10.61%	432,937	0.04%
P. abelii	PA_A952	235,278,445	22.17%	71,714,221	6.76%	478,242	0.05%
P. abelii	PA_A953	165,539,909	19.16%	58,112,480	6.73%	317,502	0.04%
P. abelii	PA_A955	157,826,898	13.71%	74,849,426	6.50%	446,462	0.04%
P. abelii	PA_B017	19,106,249	1.71%	214,074,036	19.21%	22,808,858	2.05%
P. abelii	PA_B018	14,010,011	1.15%	186,627,308	15.38%	25,827,927	2.13%
P. abelii	PA_B019	17,930,726	1.68%	198,400,997	18.62%	12,715,277	1.19%
P. abelii	PA_B020	16,041,711	1.51%	165,324,329	15.54%	17,524,242	1.65%
P. abelii	PA_KB4361	26,047,292	5.18%	70,810,607	14.09%	1,181	0.00%
P. abelii	PA_KB4661	14,107,709	3.57%	57,371,308	14.52%	2,453	0.00%
P. abelii	PA_KB5883	39,113,310	8.31%	68,810,275	14.62%	1,424	0.00%
P. abelii	PA_KB9258	28,230,836	5.17%	83,544,289	15.29%	1,428	0.00%
P. abelii	PA_SB550	15,906,300	3.78%	64,792,367	15.39%	3,201	0.00%
P. pygmaeus	PP_5062	7,891,952	1.52%	51,019,044	9.80%	467,125	0.09%
P. pygmaeus	PP_A938	143,064,905	16.28%	57,302,521	6.52%	474,279	0.05%
P. pygmaeus	PP_A939	166,225,495	16.91%	69,332,393	7.05%	418,801	0.04%
P. pygmaeus	PP_A940	21,089,546	2.40%	66,213,038	7.53%	458,225	0.05%
P. pygmaeus	PP_A941	75,201,269	7.72%	65,002,932	6.67%	433,456	0.04%
P. pygmaeus	PP_A942	186,368,059	16.64%	79,902,965	7.14%	523,362	0.05%
P. pygmaeus	PP_A943	166,560,691	14.65%	81,341,070	7.15%	471,098	0.04%
P. pygmaeus	PP_A944	163,330,781	14.71%	86,536,463	7.79%	511,083	0.05%
P. pygmaeus	PP_A946	67,691,610	7.17%	78,230,815	8.28%	389,434	0.04%
P. pygmaeus	PP_A983	12,085,031	1.05%	130,143,364	11.31%	1,089,473	0.09%
P. pygmaeus	PP_A984	13,694,397	1.17%	133,233,733	11.43%	1,596,078	0.14%
P. pygmaeus	PP_A985	10,739,096	0.90%	139,051,397	11.70%	1,577,301	0.13%
P. pygmaeus	PP_A987	12,414,888	1.05%	123,045,423	10.41%	1,325,398	0.11%
P. pygmaeus	PP_A988	12,024,936	1.02%	117,985,391	9.96%	1,155,559	0.10%
P. pygmaeus	PP_A989	12,366,091	1.05%	123,005,953	10.40%	7,203,544	0.61%
P. pygmaeus	PP_KB4204	7,718,787	1.58%	78,293,575	16.03%	1,510	0.00%
P. pygmaeus	PP_KB5404	25,789,966	2.11%	227,395,875	18.59%	34,375	0.00%
P. pygmaeus	PP_KB5405	35,353,666	7.84%	63,732,088	14.14%	55,912	0.01%
P. pygmaeus	PP_KB5406	12,921,773	3.02%	61,347,330	14.35%	1,655	0.00%
P. pygmaeus	PP_KB5543	52,266,887	9.83%	72,854,176	13.71%	16,913	0.00%

[a]reads whose mate mapped to a different contig; [b]reads with non-unique mapping

Supporting Table S3. Overview of effective sequence coverage distribution[a]. For each study individual, we list the percentage of bases covered by at least 2, 5, 7, 10, and 15 unique reads.

Individual ID	Mean sequencing depth[a]	%bases≥2x	%bases≥5x	%bases≥7x	%bases≥10x	%bases≥15x
PA_A947	27.39	93.90	91.70	90.20	87.60	82.20
PA_A948	23.71	93.80	91.00	89.00	85.60	77.60
PA_A949	27.39	93.90	91.60	90.10	87.50	82.00
PA_A950	26.28	93.80	90.90	88.90	85.50	78.10
PA_A952	21.03	93.40	90.40	88.20	84.00	73.90
PA_A953	17.78	92.50	88.50	85.40	79.30	64.60
PA_A955	25.27	93.60	90.90	89.00	85.70	78.30
PA_B017	13.74	90.50	75.10	61.70	44.10	27.30
PA_B018	16.31	92.60	85.30	77.60	62.60	39.50
PA_B019	16.92	91.60	79.00	67.60	52.40	37.10
PA_B020	16.30	93.20	87.30	80.10	63.70	37.40
PA_KB4361	5.66	84.90	61.90	39.40	13.00	1.00
PA_KB4661	4.76	82.30	51.20	27.30	6.60	0.40
PA_KB5883	4.99	82.00	52.60	30.50	9.20	0.80
PA_KB9258	5.79	84.60	63.20	41.30	14.20	1.10
PA_SB550	4.86	83.20	52.90	28.20	6.70	0.50
PP_5062	13.81	90.70	81.10	72.80	60.60	43.20
PP_A938	18.62	91.90	87.40	83.80	77.30	62.90
PP_A939	20.48	92.30	89.10	86.60	82.00	70.70
PP_A940	21.80	93.00	90.00	87.90	84.10	74.60
PP_A941	23.17	92.20	88.90	86.60	82.40	73.30
PP_A942	23.12	92.70	90.00	88.20	85.00	77.30
PP_A943	24.17	92.80	90.30	88.70	86.00	79.20
PP_A944	23.32	92.80	90.30	88.50	85.20	76.80
PP_A946	22.39	90.50	84.70	80.90	75.10	64.80
PP_A983	29.71	94.30	92.20	91.00	89.30	85.40
PP_A984	29.89	94.20	92.20	91.00	89.40	86.70
PP_A985	30.13	94.40	92.40	91.20	89.50	85.80
PP_A987	30.65	94.30	92.30	91.20	89.60	86.90
PP_A988	31.06	94.30	92.30	91.20	89.50	86.00
PP_A989	27.30	94.20	91.90	90.70	88.90	85.40
PP_KB4204	5.61	84.10	60.90	38.60	12.60	1.00
PP_KB5404	12.24	90.50	85.00	80.50	68.50	35.60
PP_KB5405	5.61	84.00	61.30	39.40	13.00	0.90
PP_KB5406	4.90	82.60	52.80	28.90	7.40	0.60
PP_KB5543	6.03	83.10	60.20	41.80	19.10	3.00

[a]mean effective whole-genome sequencing coverage (estimated from the filtered BAM files, see Materials and Methods)

Chapter 5

Discordant sex-specific evolutionary histories of orangutans inferred from deep sequencing of Y chromosomes and mitochondrial genomes

Maja P. Greminger[1], Alexander Nater[1,2], Christian Roos[3], Benoit Goossens[4,5,6], Marta Gut[7], Ivo G. Gut[7], Carel P. van Schaik[1], Tomas Marques-Bonet[7,8], and Michael Krützen[1]

[1]Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

[2]Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

[3]Gene Bank of Primates and Primate Genetics Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen, Germany.

[4]Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom

[5]Danau Girang Field Centre, c/o Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[6]Sabah Wildlife Department, 88100 Kota Kinabalu, Sabah, Malaysia

[7]Centro Nacional de Análisis Genómico, Barcelona, Spain

[8]CREA, Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain

5.1 Abstract

Because sex-biased dispersal may strongly impact genetic diversity and structure of populations, independently tracing female- and male-specific lineages is critical for a comprehensive understanding of evolutionary history. Yet, while the maternally inherited mitochondrial genome (mitogenome) is widely studied, large-scale 'genomic' data of the paternally transmitted Y chromosome are largely limited to humans. Here, we developed a novel bioinformatics strategy, widely applicable to other mammal species, to extract Y-specific single-copy sequences from whole-genome sequencing data. We traced both sex-specific lineages on a genomics scale in orangutans (genus: *Pongo*), the only Asian great apes. Including samples representing the entire range of extant orangutan populations, we produced >673 kilobases of Y chromosome-specific data from 13 males as well as 50 mitogenomes. The speciation process of Bornean and Sumatran orangutans was strongly impacted by extraordinary levels of male-biased dispersal and female philopatry, in combination with the complex geological and climatic history of Southeast Asian Sundaland. In sharp contrast to geographically deeply structured mitogenome lineages (~3.97 Ma coalescent time for *Pongo*), Y chromosomes showed no within-species structure and more recent coalescence (~0.43 Ma). We estimated cessation of male-mediated gene flow between species to be considerably earlier than proposed previously. The Y-chromosomal coalescent time ~430 ka implies that habitat conditions during glacials in the late Pleistocene have prevented orangutans, and probably other rainforest-dependent species, to cross the exposed landmass between Sundaland islands. Our study demonstrates the importance and power of genomic Y-specific data for studying the evolutionary history of a taxon.

5.2 Introduction

In mammals, Y-chromosomal data represent an essential complement to maternally and biparentally inherited genetic markers by providing insight into the patrilineal evolutionary history (Prugnolle & de Meeus 2002; Handley & Perrin 2007a). Nevertheless, Y-specific data have remained elusive (Chapter 2; Greminger *et al.* 2010) and phylogenetic research has mainly relied on maternally inherited mitochondrial DNA (mtDNA). However, tracing both female and male-specific lineages is crucial as sex-biased dispersal may have profound effects on the genetic diversity and structure of populations, as for instance in mammals, males often disperse much farther than females (Handley & Perrin 2007a). Sex-specific evolutionary histories can be investigated by contrasting mtDNA sequence information with paternally inherited Y chromosome polymorphisms. In species with female philopatry and male-biased dispersal, mtDNA is expected to be highly structured, while Y-chromosomal and autosomal genetic variation will be homogenized by male-mediated gene flow.

The Y chromosome has been studied most comprehensively in humans. Here it provided, along with autosomal and mtDNA variation, the backbone for our understanding of human evolutionary history (e.g. Heyer *et al.* 2012; Hughes & Rozen 2012; Poznik *et al.* 2013; Wei *et al.* 2013a). Combined analyses of mtDNA and Y-specific loci also revealed contrasting patterns of female and male gene flow in for example non-human primates (e.g. Eriksson *et al.* 2006; Douadi *et al.* 2007b; Chan *et al.* 2012), bears (Bidon *et al.* 2014), shrews (Yannic *et al.* 2012), and canids (Hailer & Leonard 2008). Yet, despite its great importance, few Y-specific data are available for most non-human mammals (Chapter 2; Greminger *et al.* 2010). Present genetic information is often limited to very short DNA sequences, few microsatellite markers, or single-nucleotide polymorphisms (SNPs) ascertained in a small number of individuals and subsequently genotyped in a larger panel (e.g. Hailer & Leonard 2008; Nater *et al.* 2011; Chan *et al.* 2012; Yannic *et al.* 2012). To our knowledge, large scale 'genomic' Y-chromosomal sequence data of hundreds of kilobases across populations are currently only available for bears (Bidon *et al.* 2014), horses (Wallner *et al.* 2013; Schubert *et al.* 2014), and humans (e.g. Hughes & Rozen 2012; Poznik *et al.* 2013; Wei *et al.* 2013a).

The main reason for this deficit lies in the highly complex architecture of the Y chromosome. Only certain regions of the Y chromosome contain truly Y-specific and single-copy (i.e. unique) DNA sequence (Skaletsky *et al.* 2003a; Bellott *et al.* 2014; Soh *et al.* 2014). Because of the repetitive and palindromic nature of the Y chromosome, only the human, chimpanzee, rhesus macaque, and mouse Y chromosomes are completely sequenced to date (Skaletsky *et al.* 2003a; Hughes *et al.* 2005; Hughes *et al.* 2010; Hughes *et al.* 2012; Soh *et al.* 2014). Recently, however, also the complete X-degenerate regions (the only useful regions for tracing paternal lineages) of marmoset, rat, bull, and opossum have become available (Bellott *et al.* 2014), with full sequences of the male-specific region of the Y chromosome (MSY) being on their way (Bellott *et al.* 2014; Soh *et al.* 2014). Even for species without a reference Y chromosome, high-throughput sequencing offers new possibilities to generate genomic MSY-

specific data. One strategy will be to use whole-genome sequencing data to filter out Y-specific sequence reads by aligning reads to a reference Y chromosome of a related species in combination with the application of specific filters considering the unique properties of the Y chromosome.

Genomic MSY sequences will provide valuable insights into the evolutionary history of orangutans (genus: *Pongo*)—the only Asian great apes. Orangutans are unique among great apes (Eriksson *et al.* 2006; Douadi *et al.* 2007b; Langergraber *et al.* 2007; Heyer *et al.* 2012) in that they exhibit strong male-biased dispersal and female philopatry, according to both behavioral (van Noordwijk *et al.* 2012) and genetic data (Arora *et al.* 2012; Nietlisbach *et al.* 2012). Thus, to understand population history and phylogeographic patterns of this genus, tracing paternal lineages is of particular importance.

Orangutans are currently endemic to the Sundaland islands of Sumatra and Borneo in Southeast Asia (Figure 1), whose flora and fauna have been severely affected by Quaternary climatic oscillations. Falling sea levels during recurring glacial periods caused temporary exposure of the continental shelf, repeatedly reconnecting the islands of Sundaland (Voris 2000) and thus potentially facilitating gene flow. In species with strong sex-biased dispersal, discrepancies in the coalescent times of male- and female-specific lineages between islands are therefore expected: coalescent times of lineages transmitted through the philopatric sex should reflect splits going back to the early Pleistocene when the Sundaland islands reached their present shape (Meijaard 2004), while lineages specific to the dispersing sex should show signals of inter-island migration as recent as the last glacial period (110–18 ka).

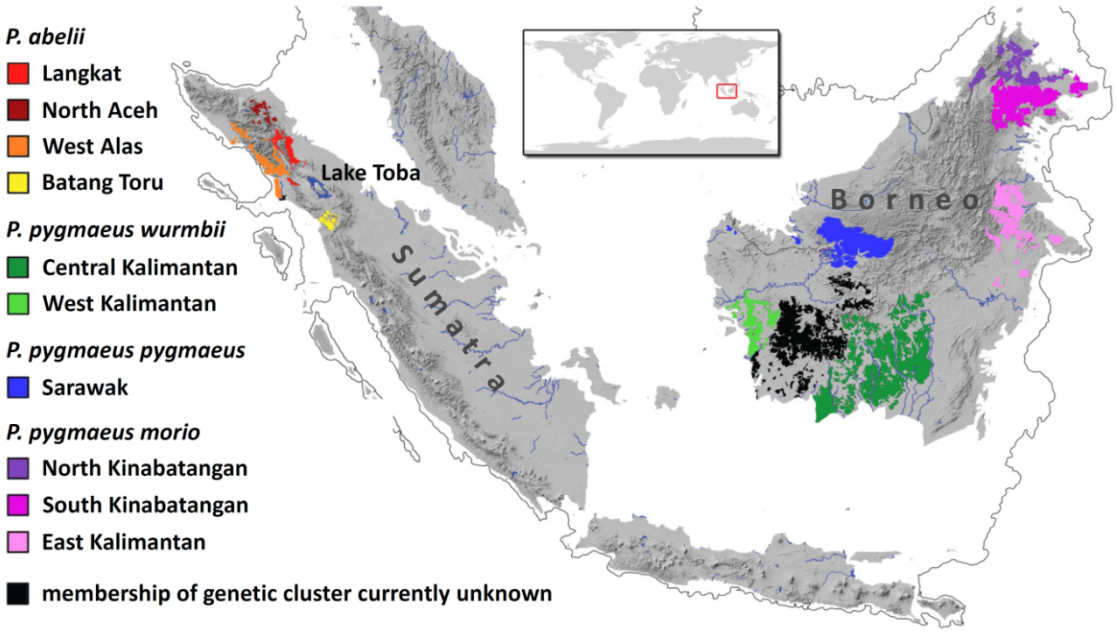


Figure 1. Current distribution of orangutans. Extant populations (i.e. major genetic clusters) are indicated by different colors. The contour of the exposed Sunda Shelf during the last glacial maximum 19–26 ka is shown by a grey line (-120 meters below current sea level).

In rainforest-dependent species such as orangutans, however, matters are more complicated because of other potentially isolating forces during the Pleistocene. A savanna corridor (Bird *et al.* 2005; Slik *et al.* 2011) combined with large river systems dissecting the exposed Sunda Shelf might have imposed significant dispersal barrier (Rijksen & Meijaard 1999; Harrison *et al.* 2006). Furthermore, cyclical climate change caused fluctuations in the Sundaland rainforest distribution, as glacial periods were considerably more arid and seasonal than inter-glacials (Morley 2000), leading to repeated isolation and reconnection of populations. Finally, the Sunda archipelago was subjected to volcanic activity, in particular of Mount Toba in northern Sumatra, which had at least four major and numerous smaller eruptions during the last 1.2 million years (Chesner *et al.* 1991).

The evolutionary history of orangutans has been strongly influenced by the Pleistocene environmental changes (e.g. Delgado & van Schaik 2000; Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2013; Nater *et al.* 2015). As predicted, Nater *et al.* (2011) found a more recent Y-chromosomal coalescence compared to mtDNA between Bornean (*P. pygmaeus*) and Sumatran (*P. abelii*) orangutans. The estimated divergence time for Y-chromosomal markers was 168 ka, suggesting recent male-mediated gene flow across the Sunda Shelf. In contrast, several studies found very old mtDNA divergence of 1–5 million years between the two currently recognized species (Zhi *et al.* 1996; Warren *et al.* 2001; Zhang *et al.* 2001; Steiper 2006; Nater *et al.* 2011). Unexpectedly, Nater *et al.* (2011) also observed that the lineage of Batang Toru—the only remaining Sumatran population south of the Toba caldera—was more closely related to Bornean mtDNA lineages than to other Sumatran orangutans.

For comparative analyses of male- and female-specific lineages, genetic data with same underlying mutation model and comparable mutation rates are required for each marker system (Poznik *et al.* 2013). Genetic data in Nater *et al.* (2011) did not fulfill these criteria, as Y-data were limited to only 11 rapidly mutating microsatellites and seven SNPs. Microsatellite markers are ill-suited to estimate coalescent times and effective population sizes (N_e) because: (i) their mutation rates range over orders of magnitude among loci (Ellegren 2004), (ii) their step-wise mutation mode leads to frequent homoplasy (Estoup *et al.* 2002), and (iii) phylogenetic trees cannot be rooted with outgroups. In line with this notion, large discrepancies have been observed between genomic MSY sequence data and microsatellite data in humans, with older time to the most recent common ancestor (T_{MRCA}) of MSY sequences compared to microsatellites (e.g. Pritchard *et al.* 1999; Poznik *et al.* 2013; Wei *et al.* 2013b).

In orangutans, genomic MSY sequence data will finally shed light into the long-lasting debate when male-mediated gene flow ceased between Borneo and Sumatra (Harrison *et al.* 2006; Kanthaswamy *et al.* 2006; Steiper 2006; Locke *et al.* 2011; Nater *et al.* 2011; Nater *et al.* 2015). To address such central questions of speciation and demography in orangutans, we generated and analyzed extensive genomic MSY sequences for thirteen orangutan males who represent almost the entire current geographic distribution of genus *Pongo* (Figure 1, Table S1). For this, we developed a widely applicable bioinformatics strategy to retrieve MSY-

specific, single-copy sequences from whole-genome sequencing data. We also produced 50 complete mitochondrial genomes ('mitogenomes'), as previous mtDNA data were restricted to short sequences, which not always reveal actual evolutionary history (Knaus *et al.* 2011).

5.3 Results

MSY sequences

To generate genomic male-specific data, we developed a novel bioinformatics strategy to extract MSY-specific sequences from whole-genome sequencing data (Figure 2; *Materials and Methods*). Due to the lack of a species-specific reference Y chromosome, we mapped previously generated Illumina whole-genome sequencing reads (Chapter 4; Locke *et al.* 2011; Prado-Martinez *et al.* 2013) of our 13 study males (Tables 1 and S1) to the reference Y of a related taxon (in our case humans with ~12-20 million years divergence; Steiper & Young 2006; Prado-Martinez *et al.* 2013). We applied several filters to ensure male-specificity and single-copy status of the generated MSY sequences. (i) We simultaneously mapped sequencing reads to the whole orangutan reference genome (*PonAbe2*, Locke *et al.* 2011) and not just the human reference Y chromosome, reducing spurious mapping of autosomal reads to the Y chromosome and allowing subsequent identification of reads that also aligned to the X or autosomal chromosomes. (ii) We exclusively accepted reads that mapped in a proper pair, i.e. where both read mates mapped to the Y chromosome, which considerably increased confidence in Y-specific mapping. (iii) We also mapped whole-genome sequencing reads of 23 orangutan females to the human Y reference chromosome and excluded all sites where reads had mapped from the male Y sequence data. (iv) To exclude potential repetitive regions, we filtered non-uniquely mapped reads as well as positions with sequence coverage greater than two times the median coverage for each individual, as extensive coverage can be indicative for repetitive regions which might appear as collapsed regions on the Y reference chromosome. (v) To ensure that we only targeted unique, single-copy MSY regions, we exclusively retained reads mapping to four well-established X-degenerate regions of the MSY in humans (Table 2, cf. Wei *et al.* 2013a).

We obtained 2,825,271 bp of MSY sequences among the 13 orangutan males from the four selected X-degenerate regions (corresponding to 3,854,654 bp in humans). As expected, individual mean MSY sequence depth was about half (average: 54.4%) of that reported for the autosomes (cf. Chapter 4), and ranged from 2.79–16.62x (Table S2). For analyses, we only kept sites without missing data, i.e. with a genotype in all study males. Because genomes of some individuals had been sequenced to only low coverage (~5-7x; Locke *et al.* 2011), this left us with 673,165 bp of MSY sequences (deposited on Genbank accession no. NC_xxxxxxx-NC_ xxxxxxx). We identified 1,317 SNPs among the 13 males (Table 3), which corresponds to a SNP density of 1 SNP every 511 bp in our dataset.

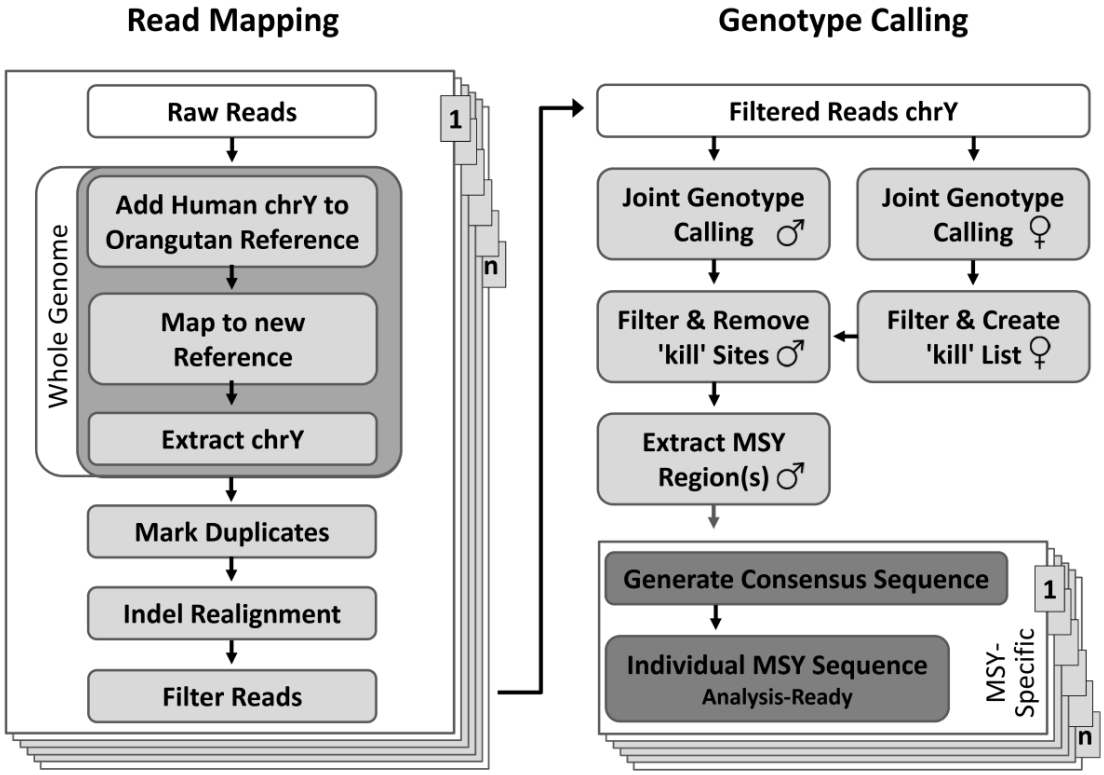


Figure 2. Bioinformatics strategy to extract MSY-specific, single-copy sequences from whole-genome sequencing data. Steps performed at the individual level are graphically illustrated as layers with numbering (1–n). Figure style adapted from the GATK guide.

As resource for further population genetic studies, we also identified classical microsatellite markers in the orangutan MSY. We found 71 microsatellites matching our search criteria (*Materials and Methods*). Forty-seven microsatellites were di-nucleotide motifs, seven tri-nucleotide and 17 tetra-nucleotide motifs (Table S3). Sequences of each microsatellite marker including 600 bp flanking sequences are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN].

Mitogenomes

We also generated data from complete mitogenomes (*Materials and Methods*) for 50 individuals (Tables 3 and S1), which spanned the entire geographic distribution of extant orangutans (available from Genbank accession no. NC_xxxxxxx-NC_ xxxxxxx). In total, we identified 1,512 SNPs among all 50 individuals (Table 3).

Table 3. Summary statistics for mitogenomes and MSY sequences. For mitogenomes, we provide statistics for Sumatran orangutans both with and without the Batang Toru (BT) population, which was found to show higher affinity to Bornean orangutans than to other Sumatran populations for mtDNA (present study, Nater *et al.* 2011).

	mtDNA							
Species	**$N_{Samples}$**	**$N_{Hapl.}$**	**N_{Sites}**	**N_{SNPs}**	**θ_π[a] ± s.d.**	**θ_W[b] ± s.d.**	**N_e [θ_π][c] ± s.d.**	**N_e [θ_W][c] ± s.d.**
Pongo	50	36	15,397	1,512	0.03512 ± 0.00159	0.02335 ± 0.00606	132,528 ± 6,000	88,113 ± 22,867
P. abelii	25	17	15,397	1,129	0.02119 ± 0.00449	0.02182 ± 0.00624	79,962 ± 16,943	82,339 ± 23,547
P. abelii excl. BT	22	14	15,397	431	0.01125 ± 0.00074	0.00784 ± 0.00255	42,452 ± 2,792	29,584 ± 9,622
P. pygmaeus	25	19	15,397	142	0.00215 ± 0.00021	0.00246 ± 0.00080	8,113 ± 792	9,283 ± 3,018
	MSY							
Species	**$N_{Samples}$**	**$N_{Hapl.}$**	**N_{Sites}**	**N_{SNPs}**	**θ_π[a] ± s.d.**	**θ_W[b] ± s.d.**	**N_e [θ_π][d] ± s.d.**	**N_e [θ_W][d] ± s.d.**
Pongo	13	13	673,165	1,317	0.00070 ± 0.00007	0.00063 ± 0.00024	28,000 ± 2,800	25,200 ± 9,600
P. abelii	5	5	673,165	422	0.00030 ± 0.00005	0.00030 ± 0.00015	12,000 ± 2,000	12,000 ± 6,000
P. pygmaeus	8	8	673,165	563	0.00027 ± 0.00003	0.00032 ± 0.00014	10,800 ± 1,200	12,800 ± 5,600

[a]Theta from π ± standard deviation

[b]Theta from the number of segregating sites ± standard deviation

[c]long-term effective population size, calculated as $N_e=\theta_{(\pi/W)}/\mu$ assuming a mutation rate of 2.65×10^{-7}/site/generation with a generation time of 25 years

[d]long-term effective population size, calculated as $N_e=\theta_{(\pi/W)}/\mu$ assuming a mutation rate of 2.50×10^{-8}/site/generation with a generation time of 25 years

Sex-specific phylogenies and molecular dating

Bayesian phylogenetic analyses of the MSY and mitogenomes revealed strict separation (posterior probabilities 1.00) of Sumatran and Bornean orangutans for both marker systems (Figure 3). Yet, matrilineal and patrilineal phylogenies differed fundamentally.

The rooted mitogenome tree exhibited clear geographical structure (Figure 3a). We estimated the T_{MRCA} of all extant orangutan mtDNA lineages to be ~3.97 Ma (95% highest posterior density interval: 2.35–5.57 Ma). Sumatran orangutans formed a paraphyletic group, with Batang Toru being more closely related to the Bornean lineage than to the other Sumatran populations. The lineage of Batang Toru and that giving rise to extant Bornean orangutans separated ~2.41 Ma (1.26–3.42 Ma). In contrast to Sumatran orangutans, Bornean populations formed a monophyletic group with recent mtDNA coalescence ~160 ka (94–227 ka). The two orangutan populations north and south of the Kinabatangan river in Northern Borneo (North and South Kinabatangan) were basal to all other Bornean populations and diverged rather recently from each other (~40 ka, 20–70 ka).

In sharp contrast to the strong geographic structure in the mitogenome phylogeny, the rooted MSY tree showed no obvious geographical clustering within islands, strongly suggesting male-mediated long-distance gene flow (Figure 3b). Both orangutan species were reciprocally monophyletic for the MSY, with the male from the Batang Toru population being well embedded within the Sumatran MSY phylogeny. We estimated the coalescent time of all orangutan Y chromosomes to be ~426 ka (296–576 ka), almost an order of magnitude more recent than the T_{MRCA} based on mitogenomes. Y chromosomes of Sumatran orangutans coalesced ~130 ka (87–174 ka), those of Bornean orangutans ~110 ka (81–149 ka). Detailed summary statistics of Bayesian phylogenetic analyses of MSY and mitogenomes are provided in Tables S4 and S5.

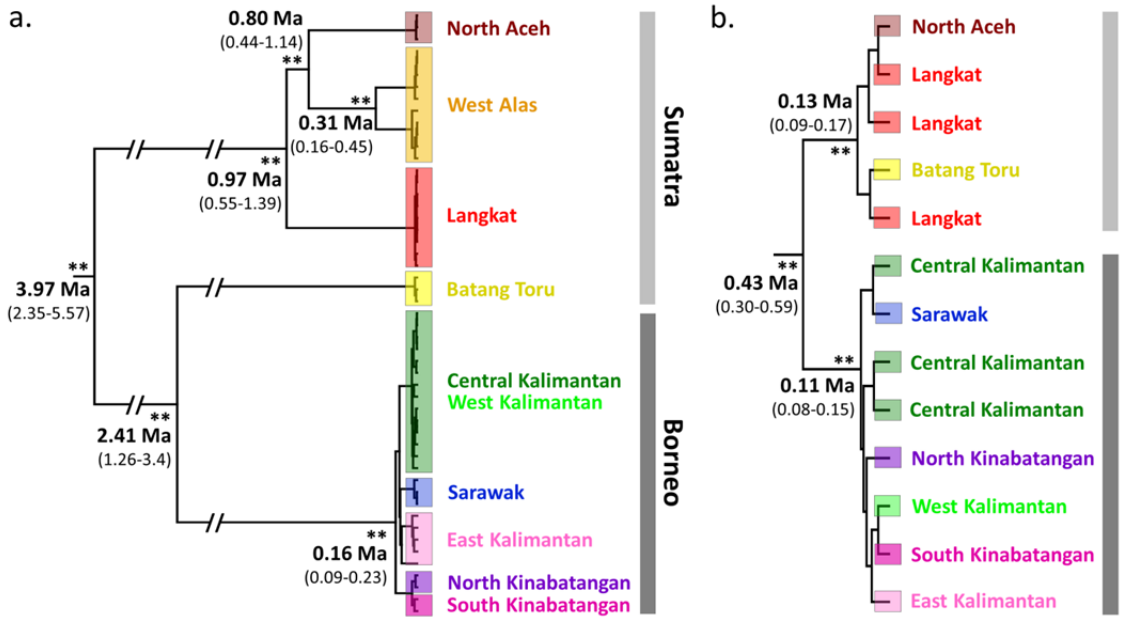


Figure 3. Bayesian phylogenetic trees for (a) mitogenomes and (b) MSY. The posterior probability of all major nodes was 1.00 (**). The node ages are mean values of the posterior probability distribution (Tables S6 and S7). The mitogenome tree is rooted with a human and a central chimpanzee sequence, the MSY tree with a human sequence (not shown).

Population differentiation and divergence

Bornean and Sumatran orangutans were significantly differentiated for both mitogenomes (Φ_{ST} = 0.797, p << 0.00001) and MSY (Φ_{ST} = 0.734, p << 0.00079). Patterns of pairwise genotype-sharing at polymorphic sites, however, were contrasting between mitogenomes and MSY (Figure 4). While mitogenome pairwise genotype-sharing was strongly geographically structured (Figure 4, above the diagonal; Figure S1), MSY estimates were much less so (Figure 4, below diagonal). For mitogenomes, we observed a clear split between Sumatran populations north of Lake Toba and Batang Toru, located south of Lake Toba. The Batang Toru male shared a higher percentage of mtDNA alleles with Bornean orangutans than with other Sumatran orangutans. This finding was confirmed for the two female Batang Toru orangutans in our dataset in a mitogenome genotype-sharing matrix including all 50 individuals (Figure S1).

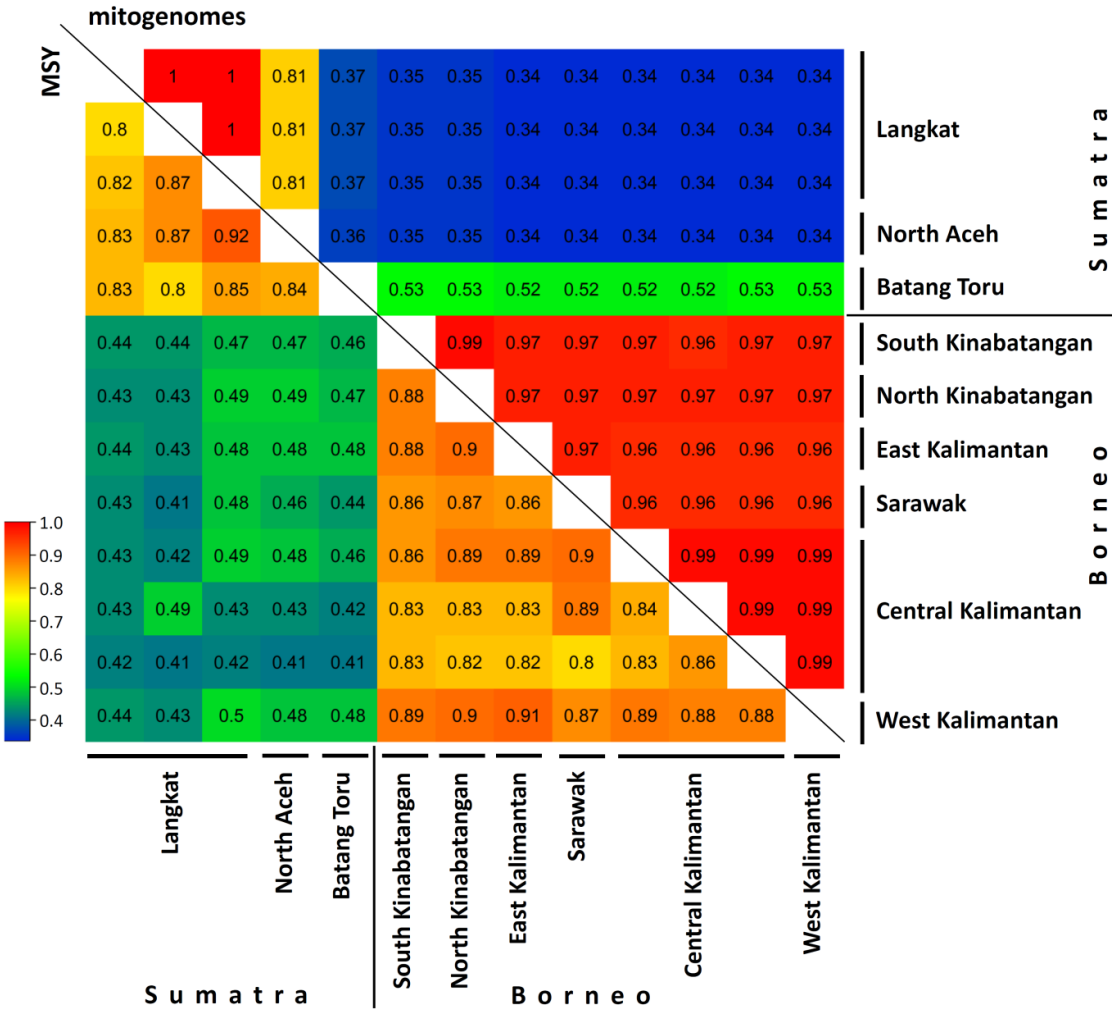


Figure 4. Genotype-sharing matrix for mitogenomes (above the diagonal) and MSY (below the diagonal) for all male orangutans. Genotype-sharing coefficients were calculated as the proportion of shared genotypes between a pair of males across all SNPs. A value of 1 indicates that two males have identical genotypes at all polymorphic sites; a value of 0 means that they have different genotypes at all SNP positions.

Genetic diversity and N_e

Mitogenomes and MSY showed different patterns of genetic diversity between Sumatran and Bornean orangutans (Table 3). For mitogenomes, nucleotide diversity (θ) was an order of magnitude larger for Sumatran orangutans compared to Borneans ($N_e[\theta_\pi]$:~0.0212 versus ~0.0022). For the MSY, however, nucleotide diversities were similar for the two species (~0.0003).

Sumatran orangutans showed a strong discrepancy in the long-term N_e estimates of each marker system (mitogenomes: ~79,962 versus MSY: ~12,000). When Batang Toru individuals were excluded from the Sumatran dataset, mitogenome N_e estimates were reduced by 47%–64% (depending on the estimator of θ, Table 1). In Borneo, long-term N_e estimates were

slightly lower for mitogenomes (~8,113) than for the MSY (~10,800), although both were within the same 95% highest posterior density interval.

On the population level, mitogenome genetic diversity within populations was generally lower for Sumatran orangutans compared to Borneans (Table 4), except for the large West Alas cluster, which has by far the largest census size of all extant orangutan populations (Wich *et al.* 2008). Irrespective of the small sample sizes, we detected particularly low genetic diversity for the two Sumatran populations Langkat and North Aceh.

Table 4. Within population diversity of mitogenomes.

	$N_{Samples}$	$N_{Haplotypes}$	N_{SNPs}	π
All *Pongo*	50	36	1,512	0.03512
Sumatra	25	17	1,129	0.02119
Langkat	9	3	4	0.00007
North Aceh	3	3	2	0.00009
West Alas	10	8	102	0.00350
Batang Toru	3	3	6	0.00026
Borneo	25	19	142	0.00215
South Kinabatangan	2	2	3	0.00020
North Kinabatangan	2	2	6	0.00039
East Kalimantan	3	2	7	0.00030
Sarawak	4	4	22	0.00072
Central/West Kalimantan	14	9	32	0.00050

π: nucleotide diversity

Analysis of molecular variance of mitogenomes

The AMOVA of mitogenomes supported the differentiation between the two islands, with 72% of the total molecular variance being partitioned between Bornean and Sumatran orangutans (Table 5). Defining Batang Toru as a third group, the among-group variance even increased to 84%. At the species level, AMOVA results confirmed the much higher within-population diversity of Bornean orangutans compared with Sumatrans. In Borneo, 17% of the total variation was found within population whereas for Sumatran populations it was only 0.86%.

Table 5. Results of AMOVA of mitogenomes

	Variance components	% Variance explained
Groups: Sumatra and Borneo		
Among groups/species	321.29*	72.37
Among populations within species	119.80*	26.98
Within populations	2.88*	0.65
Groups: Sumatra, Batang Toru and Borneo		
Among groups	372.65*	84.33
Among populations within groups	66.37*	15.02
Within populations	2.88*	0.65
Sumatra		
Among populations	203.51*	99.14
Within populations	1.76*	0.86
Borneo		
Among populations	19.19*	82.75
Within populations	4.00*	17.25

*$p < 0.01$

5.4 Discussion

We simultaneously traced both the male- and female-specific evolutionary history on a genomic level in orangutans. Analyzing a unique dataset encompassing orangutans of almost all extant populations, we found that population history and phylogeography of females and males differed drastically. While mitogenome lineages showed deep geographic structure and old coalescent times, MSY haplotypes were not geographically structured within species and coalesced much more recently.

Speciation and gene flow

Y chromosomes of all extant orangutans coalesced ~430 ka, in stark contrast to the coalescence of mtDNA lineages of both species at 2.37 Ma. Thus, the speciation process of Bornean and Sumatran orangutans had been strongly impacted by the extraordinary high levels of female philopatry and male-biased dispersal. Our data show that after the initial separation of Bornean and Sumatran orangutans in the early Pleistocene, their autosomal gene pools remained connected via strictly male-driven gene flow.

The Y-chromosomal T_{MRCA} of ~430 ka implies that male migration between the islands was possible during at least some glacial periods in the early and middle Pleistocene. However, geological and habitat conditions during glacials in the late Pleistocene seem to have prevented male orangutans, and most likely also other rainforest-dependent species, to cross the exposed Sunda Shelf. A savanna corridor (Bird *et al.* 2005), large river systems dissecting the exposed shelf (Rijksen & Meijaard 1999; Voris 2000; Harrison *et al.* 2006), or most likely both, may have imposed impassable dispersal barriers. Our results strongly contrast with previous studies (Muir *et al.* 2000; Verschoor *et al.* 2004; Kanthaswamy *et al.* 2006; Steiper 2006; Nater *et al.* 2011; Nater *et al.* 2015; but see Locke *et al.* 2011 and Mailund *et al.* 2012), which had suggested much more recent gene flow between islands. The timing of the cessation—or at least substantial reduction—of male migration between Borneo and Sumatra coincides with the beginning of the Marine Isotope Stage (MIS) 11. MIS 11 was the longest interglacial period isolating the islands in half a million years, spanning ~424–374 ka (de Vernal & Hillaire-Marcel 2008).

Population history of Bornean orangutans

Climatic and thus rainforest cover oscillations strongly impacted the population histories of the two orangutan species. Both mitogenome and MSY provide strong evidence for a major bottleneck of Bornean orangutans. In stark contrast to the deep mtDNA splits within Sumatra, Bornean mitogenome lineages coalesced ~160 ka (94–227 ka), showing that populations diverged rather recently. The T_{MRCA} for Bornean Y chromosomes was ~110 ka (81–149 ka), and thus within the range of mitogenomes. This pattern is congruent with a common population refugium during the penultimate glacial period (130–190 ka), when rainforest coverage on Borneo was drastically reduced (Morley 2000; Gathorne-Hardy *et al.* 2002; Bird

et al. 2005). Bornean orangutans also exhibit a much higher relative within-population mtDNA diversity compared to Sumatrans (17% compared to 0.86% of the total molecular variance within species; similar to Nater *et al.* 2011), providing independent support for the late-Pleistocene refugium hypothesis (Arora *et al.* 2010). In agreement with the basal position of the two Sabah populations to all other Borneans in our mitogenome phylogeny, such a common refugium was most likely located in the Crocker mountain range in northern Borneo, as proposed for the Sabah population (North and South Kinabatangan, Jalil *et al.* 2008) and other Bornean species (Barkman & Simpson 2001; Gathorne-Hardy *et al.* 2002; Cannon & Manos 2003; Quek *et al.* 2007).

Population history of Sumatran orangutans

In contrast to Bornean orangutans, the deep mitogenome splits of Sumatran populations suggest that population structure was remarkably stable for a long period of time, probably best explained by the differences in geology and environmental conditions across Sundaland. For large parts of Sumatra, rain-fall rates during glacial periods were considerably higher compared to Borneo (Gathorne-Hardy *et al.* 2002). Thus, multiple rainforest refugia likely existed in most of northern Sumatra and at the base of the Barisan Mountains (Gathorne-Hardy *et al.* 2002), which may have contributed to the deep population structure of Sumatran orangutans observed today.

Our results also support a strong impact of the Toba volcano (Chesner *et al.* 1991) on the evolutionary history of orangutans (Nater *et al.* 2011; Nater *et al.* 2015). That we observed the deepest split in the mitogenome phylogeny (~3.97 Ma) between Sumatran populations north of Lake Toba and Batang Toru to the south of it, rather than between both currently recognized orangutan species, confirms previous findings (Nater *et al.* 2011). The results also imply that the Toba region has been subject to major volcanic activity for much longer than the first recorded eruptions at ~1.2 Ma. Such an extremely old separation of mtDNA lineages within a single species is exceptional among primates (Finstermeier *et al.* 2013). The Batang Toru population likely constitutes a remnant of a large gene pool south of Lake Toba (Nater *et al.* 2011), from where orangutans colonized the island of Borneo.

Species-specific forces acting on the Y chromosome

While N_e of the MSY and mitogenomes were similar for Bornean orangutans, N_e was considerably larger for mitogenomes than for the MSY in Sumatra. In light of the more stable population history of Sumatran orangutans, one might expect higher long-term N_e of Sumatran MSY. This implies that different forces may act on the MSY in both orangutan species. First, male reproductive skew might reduce N_e and therefore T_{MRCA} of Sumatran Y chromosomes compared to Bornean orangutans. Both species strongly differ in their ecology as well as social- and mating systems (Delgado & van Schaik 2000; Wich *et al.* 2009b; Dunkel *et al.* 2013). There is increasing evidence for extensive reproductive skew among Sumatran males, whereas reproduction is more evenly distributed in Borneo (Goossens *et al.* 2006b;

Lenzi 2014). Second, positive selection acting on beneficial mutations, potentially linked to reproductive skew, could also have reduced genetic diversity of Sumatran MSY. Third, the effective male-gene flow rate might be higher in Sumatran orangutans because males can cross rivers more easily at their headwaters since suitable habitat is provided to higher altitudes than on Borneo due to the *Massenerhebung* effect caused by the Barisan Mountain ridge (van Schaik *et al.* 1995; Rijksen & Meijaard 1999).

MSY bioinformatics strategy

Overall, our novel bioinformatics strategy proved to be extremely useful to extract MSY-specific, single-copy sequences from whole-genome sequencing data in orangutans. This was achieved through the implementation of several filtering steps dealing with the specific properties of the Y chromosome. We produced over 673 kb MSY sequence (2.83 Mb including missing genotypes) for male representatives of almost all extant orangutan populations and describe 1,317 novel MSY-SNPs as well as 71 MSY-microsatellite markers for *Pongo*. To our knowledge, comparable Y chromosome sequencing in non-human mammals has only been achieved for horses (Wallner *et al.* 2013; Schubert *et al.* 2014) as well as polar and brown bears (Bidon *et al.* 2014). Our results strongly contrast with a previous study based on 18 Y-linked loci, mostly microsatellite markers (Nater *et al.* 2011), which demonstrates the importance of producing comparable genetic data for the MSY (i.e. sequence data) as for other marker systems such as mitogenomes.

We expect the principle of our bioinformatics strategy to be applicable to mammalian species in general (Figure S2). Like for most mammals, there is currently no reference Y chromosome for orangutans available. Therefore we had to rely on a reference assembly of a related species (i.e. humans) for sequence read mapping. Despite the ~18 million years divergence between humans and orangutans (Steiper & Young 2006; Prado-Martinez *et al.* 2013), we obtained a high number of MSY sequences. The impact of varying Y chromosome structure among species (Bellott *et al.* 2014; Soh *et al.* 2014) on sequence read mappability might have been reduced because we exclusively targeted X-degenerate regions. Hughes *et al.* (2010) showed for human and chimpanzees that although less than 50% of ampliconic sequences have a homologous counterpart in the other species, over 90% of the X-degenerate sequences hold such a counterpart.

Our study highlights the importance of the Y chromosome for a comprehensive understanding of a species' demographic history and phylogeography. Independently tracing maternal and paternal lineages, we found that orangutan evolutionary history is not only a tale of two islands, but also one of two sexes. Males and females exhibited strikingly distinct histories—as if they were two different species.

5.5 Materials and Methods

Sampling schema

We used Illumina whole-genome short read sequence data from 36 orangutans (13 males, 23 females) representing the entire extant distribution of orangutans (Figure 1, Tables 1 and S1). Data were generated previously by Locke *et al.* (2011, n=10), Prado-Martinez *et al.* (2013, n=10), and Chapter 4 (n=16). Population provenance was genetically verified based on both mtDNA and autosomal loci as described in Chapter 4. We supplemented our data set with 13 Sanger-sequenced mitogenomes (Table S2).

Table 1. Number of study individuals per extant orangutan population. The number of individuals for which mitogenomes were obtained using Sanger sequencing is given in parenthesis.

Species	Population	$N_{Mitogenomes}$	N_{MSY}
P. abelii	Langkat (LK)	9 (2)	3
P. abelii	North Aceh (NA)	3 (1)	1
P. abelii	West Alas (WA)	10 (3)	0
P. abelii	Batang Toru (BT)	3 (1)	1
P. pygmaeus	South Kinabatangan (SK)	2	1
P. pygmaeus	North Kinabatangan (NK)	2	1
P. pygmaeus	East Kalimantan (EK)	3 (1)	1
P. pygmaeus	Sarawak (SR)	4	1
P. pygmaeus	Central Kalimantan (CK)	13 (5)	3
P. pygmaeus	West Kalimantan (WK)	1	1
Total		50	13

MSY bioinformatics strategy

Only certain regions of the Y chromosome contain truly Y-specific and single-copied (i.e. unique) DNA sequence, thus are of use for population genetic analyses. For example, the X and Y sex chromosomes share widespread sequence homologies because they evolved from an ordinary pair of autosomes (Lahn & Page 1999; Skaletsky *et al.* 2003a). In addition, retrotransposition and gene conversion led to the incorporation of autosomal sequences into the Y chromosome and vice versa (Steinemann & Steinemann 1992; Skaletsky *et al.* 2003a; Handley & Perrin 2006). The euchromatic DNA of the male-specific region of the Y chromosome (MSY) is mainly comprised of two sequence classes: ampliconic segments and X-degenerate (also called 'ancestral single-copy') sequences (Skaletsky *et al.* 2003a; Bellott *et al.* 2014; Soh *et al.* 2014). The ampliconic regions are composed of large, nearly identical repeat units, most often arrayed as palindromes containing multi-copy gene families (Skaletsky *et al.* 2003a; Hughes & Rozen 2012), making them ill-suited for population genetic analyses. Exclusively polymorphism derived from X-degenerated sequences fulfill the criteria for population genetic analyses, i.e. Y-specificity and single-copy status (Chapter 2; Greminger *et al.* 2010). These X-degenerate sequences represent the ancestral portions of the Y chromosomes and comprise few single-copy gene or pseudogene homologues of X-linked genes (Skaletsky *et al.* 2003a; Soh *et al.* 2014).

We developed a bioinformatics strategy to extract MSY-specific, single-copy sequences from whole-genome sequencing data (Figure 2). The strategy consisted of the following steps, which ensured male-specificity and single-copy status. First, we created a new reference sequence (*PonAbe2_humanY*) by manually adding the human reference Y chromosome *(GRCh37, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/*) to the orangutan reference genome *PonAbe2* (Locke *et al.* 2011). We then used the Burrows-Wheeler Aligner v0.7.5. (BWA-MEM, Li & Durbin 2009) to map Illumina whole-genome short reads from 36 orangutans (13 males and 23 females, Tables 1 and S1) to this new reference sequence. We mapped reads for each individual separately in paired-end mode and with default settings. To reduce output file size, we removed unmapped reads on the fly using SAMtools v0.1.19. (Li *et al.* 2009). Picard v1.101 (http://picard.sourceforge.net/) was used to add read groups, sort reads, and convert sequence alignment/map (SAM) files to binary alignment/map (BAM) files.

(i) ./bwa mem -M PonAbe2_humanY input_FASTQ_F input_FASTQ_R | ./samtools view -S -F 4 -h - > output.sam

(ii) java -jar picard/AddOrReplaceReadGroups.jar INPUT=output.sam OUTPUT=output.bam SORT_ORDER=coordinate RGID=$FILEID RGLB=$LIB RGPL=ILLUMINA RGPU=$FILEID RGSM=$INDEX

(iii) java -jar picard/BuildBamIndex.jar INPUT= output.bam

We then extracted all reads which mapped to the Y chromosome using SAMtools and marked read duplicates with Picard. The Genome Analysis Toolkit v2.8.1. (GATK, McKenna *et al.* 2010;

DePristo *et al.* 2011) was used to perform local realignment around indels and filtered out duplicated reads, bad read mates, reads with mapping quality zero and reads which mapped ambiguously.

(iv) ./samtools view -h -b output.bam chrY > output_chrY.bam

(v) java -jar picard/BuildBamIndex.jar INPUT=output_chrY.bam

(vi) java -jar picard/MarkDuplicates.jar INPUT=output_chrY.bam OUTPUT=output_chrY_marked.bam METRICS_FILE=metrics.txt

(vii) java -jar picard/BuildBamIndex.jar INPUT=output_chrY_marked.bam

(viii) java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R PonAbe2_humanY -I output_chrY_marked.bam -o intervals.bed

(ix) java -jar GenomeAnalysisTK.jar -T IndelRealigner -R PonAbe2_humanY -I output_chrY_marked.bam -o output_chrY_marked_realigned.bam -targetIntervals intervals.bed

(x) java -jar GenomeAnalysisTK.jar -T PrintReads --read_filter UnmappedRead --read_filter BadMate --read_filter NotPrimaryAlignment --read_filter FailsVendorQualityCheck --read_filter MappingQualityZero --read_filter DuplicateRead -R PonAbe2_humanY -I output_chrY_marked_realigned.bam -o output_chrY_filtered.bam -log stats.txt

Genotypes at all sequenced sites were called with the *Unified Genotyper* of the GATK, applying the output mode 'EMIT_ALL_CONFIDENT_SITES'. We called genotypes in multi-sample mode (females and males separately, sample-ploidy was set to 1), producing one genomic VCF (gVCF) file for each sex. We only accepted bases/reads for genotype calling if they had Phred quality scores ≥ 30.

(xi) java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R PonAbe2_humanY --sample_ploidy 1 -mbq 30 -mmq 30 -I bamlist_M/F -o gVCF_M/F -glm SNP -gt_mode DISCOVERY -out_mode EMIT_ALL_CONFIDENT_SITES

From the gVCF file of the females, we generated a 'kill' list with the coordinates of all sites with coverage in more than one female (minimal sequence depth 2x), as these sites most likely were located in pseudoautosomal or ampliconic regions, i.e. share similarity with the X or autosomal chromosomes. To ensure Y-specificity, we therefore removed all sites of the 'kill' list from the gVCF file of the males with VCFtools v0.1.12b. (Danecek *et al.* 2011).

(xii) ./vcftools --vcf gVCF_F.vcf --recode --out gVCF_F2 --minDP 2

(xiii) sed '1,89d' gVCF_F2.vcf | awk -F $'\t' 'BEGIN {OFS = FS} {print $1, $2}' > kill.txt

(xiv) ./vcftools --vcf gVCF_M.vcf --recode --out gVCF_M2 --exclude-positions kill.txt

Finally, we used the GATK to extract sequences of four well-established X-degenerate regions of the MSY in humans (Table 2, Wei *et al.* 2013a). To be conservative, we chose regions which were longer than 1 Mb in humans and disregarded the first and last 300 kb of each region to account for potential uncertainties regarding region boundaries, leaving us with 3,854,654 bp

in total (Table 2). We exclusively retained genotype calls which were covered by minimal two reads and maximum twice the individual mean coverage. For subsequent analyses, we only kept sites for which all study males had a genotype call. For each male, we created a FASTA sequence using a custom java program (available upon request).

(xv) java -jar GenomeAnalysisTK.jar -T SelectVariants -R PonAbe2_humanY.fasta --variant gVCF_M2.vcf -o gVCF_M3.vcf -L chrY:14170438-15795786 -L chrY:16470614-17686473 -L chrY:18837846-19267356 -L chrY:21332221-21916158

(xvi) ./vcftools --vcf gVCF_M3.vcf --recode --out gVCF_M4 --minDP 2 --maxDP [2xMedianCov] --max-missing 1

Table 2. Genomic coordinates of the four extracted X-degenerate regions on the human reference Y chromosome (assembly *GRCh37/hg19*).

Region Number	Start Position	End Position	Size [bp]
1	14,170,438	15,795,786	1,625,348
2	16,470,614	17,686,473	1,215,859
3	18,837,846	19,267,356	429,510
4	21,332,221	21,916,158	583,937

MSY microsatellites

In addition, as resource for future studies, we identified male-specific microsatellite markers. For this, we created new consensus sequences for the four X-degenerate regions (Table 2) from the male with the highest mean sequence coverage (PP_A988). We required all sites to be covered at least two times and denoted all gaps to the human reference Y chromosome with 'N' to account for un-sequenced positions. We identified microsatellite motifs using the software msatcommander v0.8.2. (Faircloth 2008) applying the following settings: dinucleotide motifs with minimal eight repeats, trinucleotide motifs with minimal six repeats, and tetranucleotide motifs with minimal five repeats. Finally, we used BEDtools v2.17.0. (Quinlan & Hall 2010) to extract sequences 600 bp down- and upstream of the identified microsatellites from the consensus sequence to facilitate the design of PCR primers.

Mitogenome data

We also produced complete mitogenome sequences for all study individuals (males and females). We first created a consensus reference sequence from 13 Sanger-sequenced mitogenomes representing almost all major genetic clusters of extant orangutans using BioEdit v7.2.0. (Hall 1999). The Sanger-sequenced mitogenomes were generated via 19 PCRs with product sizes of 1.0–1.2 kb and an overlap of 100–300 bp following described methods (Roos *et al.* 2011). PCR conditions for all amplifications were identical and comprised a pre-denaturation step at 94°C for 2 minutes, followed by 40 cycles each with denaturation at

94°C for 1 minute, annealing at 52°C for 1 minute, and extension at 72°C for 1.5 minutes. At the end, a final extension step at 72°C for 5 minutes was added. PCR products were checked on 1% agarose gels, excised form the gel and after purification with the Qiagen Gel Extraction Kit, sequenced on an ABI 3130xL sequencer using the BigDye Terminator Cycle Sequencing kit (Applied Biosystems) in both directions using the amplification primers. Information on primers is available upon request.

We mapped Illumina whole-genome sequencing reads to the consensus mitochondrial reference sequence using NovoAlign v3.02. (NovoCraft), which can accurately handle reference sequences with ambiguous bases. This procedure prevented biased short read mapping due to common population-specific mutations. For each individual, we generated a FASTA sequence for the mitogenome with the *mpileup* pipeline of SAMtools. We only considered bases with both mapping and base Phred quality scores ≥ 30 and required all position to be covered between 100 and 2000 times. Finally, we visually checked the sequence alignment of all individuals in BioEdit and manually removed indels and poorly aligned positions as well as excluded the D-loop to account for sequencing and alignment errors in those regions which might inflate estimates of mtDNA diversity.

We thoroughly investigated the literature for the potential occurrence of nuclear insertions of mtDNA (numts) in the genus *Pongo*, given that this has been a concern in closely related gorillas (*Gorilla* spp.) (Thalmann *et al.* 2005). There was no indication of numts in the genus *Pongo*, which is in line with our own previous observations (Arora *et al.* 2010; Nater *et al.* 2011; Nater *et al.* 2013). Numts also seem unlikely given our high minimal sequence depth threshold.

Phylogenetic analyses

We constructed phylogenetic trees and estimated divergence dates for mitogenome and MSY sequences using the Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST v1.8.0. (Drummond *et al.* 2012). To determine the most suitable nucleotide substitution model, we conducted model selection with jModelTest v2.1.4. (Darriba *et al.* 2012). Based on the Akaike information criterion (AIC) and corrected AIC, we selected the GTR+I substitution model (Tavaré 1986) for mitogenomes and the TVM+I+G model (Posada 2003) for MSY sequences.

The mitogenome tree was rooted with a human and a central chimpanzee sequence from GenBank (accession numbers: GQ983109.1 and HN068590.1), the MSY tree with the human reference sequence *hg19*. We estimated divergence dates under a relaxed molecular clock model with uncorrelated lognormally distributed branch-specific substitution rates (Drummond *et al.* 2006). The prior distribution of node ages was generated under a birth-death speciation process (Yang & Rannala 2006). We used fossil based divergence estimates to calibrate the molecular clock by defining a normal prior distribution for certain node ages. For mitogenomes, we applied two calibration points, i.e. the *Pan-Homo* divergence with a

mean age of 6.5 Ma and a standard deviation of 0.3 (Brunet *et al.* 2002; Vignaud *et al.* 2002) and the Ponginae-Homininae divergence with a mean age of 18.3 Ma and a larger standard deviation of 3.0 (Steiper & Young 2006), which accounts for the uncertainty in the divergence date (Raaum *et al.* 2005). For MSY sequences, we used the Ponginae-Homininae divergence for calibration. We performed four independent BEAST runs for 30 million generations each for mitogenomes, with parameter sampling every 1,000 generations, and for 200 million generations each with parameter sampling every 2,000 generations for MSY sequences. We used Tracer v1.6. (Rambaut *et al.* 2013) to examine run convergence, and aimed for an effective sample size of at least 1000 for all parameters. We discarded the first 20% of samples as burn-in and combined the remaining samples of each run with LogCombiner v1.8.0. (Drummond *et al.* 2012). Maximum clade credibility trees were drawn with TreeAnnotator v1.8.0. (Drummond *et al.* 2012) and trees visualized in FigTree v1.4.0. (Rambaut 2012) and MEGA v6.06. (Tamura *et al.* 2013).

Statistical analyses

To infer the long-term N_e of the two orangutan species for both mitogenomes and MSY, we computed the estimators θ_π (based on the mean pairwise genetic distance between sequences; Nei 1987) and θ_W (based on the number of segregating sites; Watterson 1975) using DNAsp v5. (Librado & Rozas 2009). N_e was calculated from the equation Theta (θ) = $N_e\mu$, where μ is the per base mutation rate per generation (Tajima 1993), assuming a generation time of 25 years (Wich *et al.* 2009a). For mitogenomes, we used a mutation rate of 2.65×10^{-7} per site per generation as inferred with the software BEAST (Drummond *et al.* 2012) under a relaxed molecular clock model (see *Phylogenetic analyses*). This estimate was similar to the rate of 2.38×10^{-7} found by Nater *et al.* (2015). We used this procedure to account for the high variation of mtDNA mutation rates across mammals and thus to circumvent the corresponding uncertainties in mutation rate estimates (see Nabholz *et al.* 2008). For the MSY we applied a mutation rate of 2.50×10^{-8} per site per generation as inferred from human deep-rooting pedigree analysis (Xue *et al.* 2009). We obtained a similar MSY mutation rate (2.75×10^{-8}) from our phylogenetic analysis described above. The standard deviation of the θ estimates were derived from the square root of the variance assuming no recombination (Nei 1987; Tajima 1993).

To investigate pairwise sequence identity among individuals, we calculated genotype-sharing coefficients, which is the proportion of shared alleles between a pair of individuals across all SNPs. Coefficients were calculated from the MSY and mitogenome sequence alignments with MEGA v6.06. (Tamura *et al.* 2013) applying the p-distance method for pairwise distance computation. We visualized genotype-sharing coefficient matrices as heatmaps using the 'heatmap.2' function of the R package 'gplots' and customized plots with the R package 'RColorBrewer'.

Furthermore, we calculated population differentiation (Φ_{ST}) statistics between Bornean and Sumatran orangutans for mitogenomes and MSY using Arlequin v3.5.1.2. (Excoffier & Lischer

2010). For mitogenomes, we additionally performed an analysis of molecular variance (AMOVA) using Arlequin.

Acknowledgments

We thank the following institutions for supporting our research: Sabah Wildlife Department, Indonesian State Ministry for Research and Technology, Indonesian Institute of Sciences, Leuser International Foundation, Taman National Gunung Leuser, and Borneo Orangutan Survival Foundation. This study was financially supported by UZH University Research Priority Program (to MK), Leakey Foundation (to MPG), ERC Starting Grant (grant no. 260372 to TMB), Swiss National Science Foundation (grant no. 3100A-116848 to MK and CPvS), Forschungskredit University of Zurich (to MPG), Julius–Klaus Foundation (to MK), A.H. Schultz Foundation (to MK and MPG), and the Anthropological Institute & Museum at the University of Zurich.

Author Contributions

MPG and MK conceived the study. MPG performed research. BG, MPG, MK, CR, and CPvS provided samples. IG and MG carried out sequencing. CR and TMB contributed additional sequencing data. AN provided input to statistical analyses. MPG wrote the manuscript. MK and AN edited the manuscript. TMB, BG, CR, and CPvS reviewed the final manuscript. All authors read and approved the final manuscript.

5.6 Supporting Information

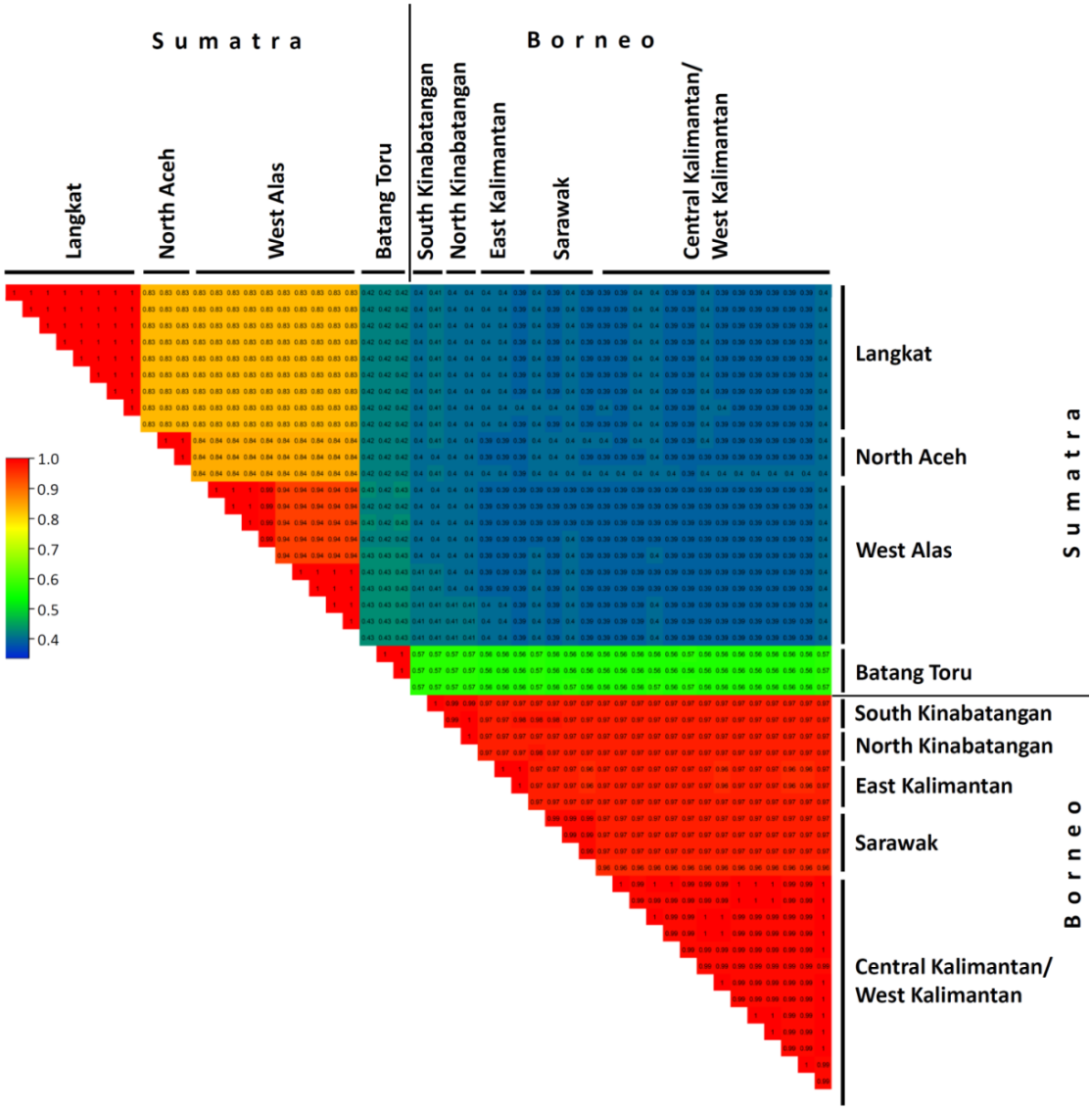


Figure S1. Genotype-sharing matrix for mitogenomes of all 50 study individuals. Genotype-sharing coefficients were calculated as the proportion of shared genotypes between a pair of individuals across all SNPs

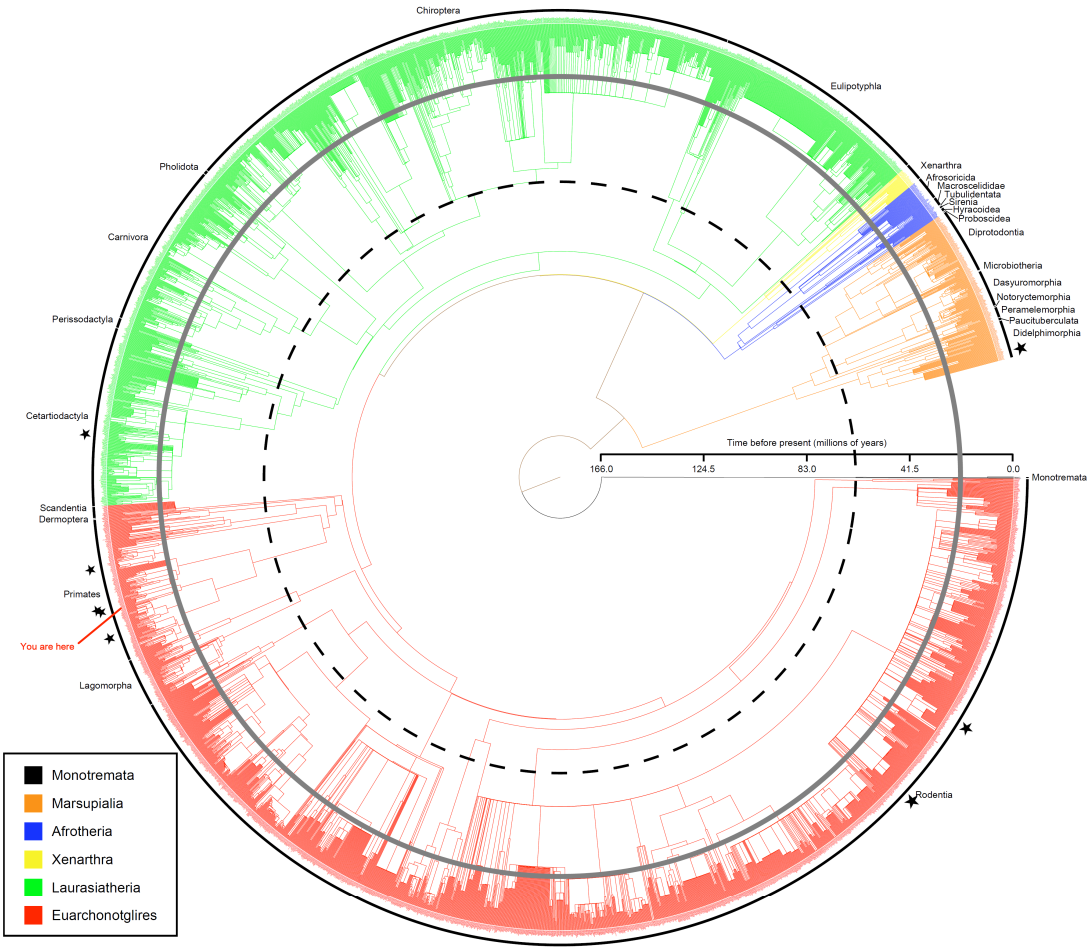


Figure S2. Tree of present-day mammals. The tick gray line indicates a divergence time of ~18 Ma as between humans and orangutans. Species for which the MSY has been completely sequenced (or at least the relevant X-degenerate regions; as at February 12, 2015) are denoted with a black asterisk. Figure adapted and reproduced with permission from Bininda-Emonds *et al.* (2007).

Table S1. Details on study individuals and available sequence data.

Species	Population	Individual ID	Individual Name	Sex	Whole-genome sequencing depth[a]	Source	Comments
P. abelii	Batang Toru	PA_B019	Afa	M	16.92	Chapter 4	Wild-born
P. abelii	Batang Toru	PA_KB9528	Baldy	F	5.79	Locke *et al.* 2011	Wild-born
P. abelii	Batang Toru	PA_BT01	BT01	M	Sanger-sequenced	This study	Wild-born
P. abelii	Langkat	PA_A947	Elsi	F	27.39	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A948	Kiki	F	23.71	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A950	Babu	F	26.28	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_A952	Buschi	M	21.03	Prado-Martinez *et al.* 2013	Wild-born
P. abelii	Langkat	PA_KB4661	Bubbles	M	4.76	Locke *et al.* 2011	Wild-born
P. abelii	Langkat	PA_KB5883	Sibu	M	4.99	Locke *et al.* 2011	Wild-born
P. abelii	Langkat	PA_518	518	M	Sanger-sequenced	This study	Wild-born
P. abelii	Langkat	PA_19	19	M	Sanger-sequenced	This study	1. Generation of wild-born Sumatra
P. abelii	Langkat	PonAbe2	Susi	F	/	Locke *et al.* 2011	Orangutan reference genome
P. abelii	North Aceh	PA_A949	Dunja	F	27.39	Prado-Martinez *et al.* 2013	1. Generation of wild-born Sumatra
P. abelii	North Aceh	PA_B018	Jeff	M	16.31	Chapter 4	Wild-born
P. abelii	North Aceh	PA_NA01	NA01	M	Sanger-sequenced	This study	Wild-born
P. abelii	West Alas	PA_A953	Vicky	F	17.78	unpubl. Prado-Martinez *et al.* 2013	Wild-born
P. abelii	West Alas	PA_A955	Suma	F	25.27	unpubl. Prado-Martinez *et al.* 2013	Wild-born
P. abelii	West Alas	PA_A964	Rochelle	F	11.06	unpubl. Prado-Martinez *et al.* 2013	1. Generation of wild-born Sumatran
P. abelii	West Alas	PA_B017	Miky	F	13.74	Chapter 4	Wild-born
P. abelii	West Alas	PA_KB4361	Likoe	F	5.66	Locke *et al.* 2011	Wild-born
P. abelii	West Alas	PA_SB550	Doris	F	4.86	Locke *et al.* 201	Wild-born
P. abelii	West Alas	PA_336	336	M	Sanger-sequenced	This study	1. Generation of wild-born Borneo
P. abelii	West Alas	PA_247	247	F	Sanger-sequenced	This study	Wild-born
P. abelii	West Alas	PA_11	11	F	Sanger-sequenced	This study	1. Generation of wild-born Borneo
P. abelii	West Alas	PA_B020	Maini	F	16.30	Chapter 4	Wild-born

Table S1 (Continued)

P. pygmaeus	Central Kalimantan	PP_A938	Lotti	F	18.62	unpubl. Prado-Martinez *et al.* 2013	1. Generation of wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_A940	Temmy	F	21.80	Prado-Martinez *et al.* 2013	1. Generation of wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_A941	Sari	F	23.17	Prado-Martinez *et al.* 2013	1. Generation of wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_A943	Tilda	F	24.17	Prado-Martinez *et al.* 2013	Wild-born
P. pygmaeus	Central Kalimantan	PP_A944	Napoleon	M	23.32	Prado-Martinez *et al.* 2013	Wild-born
P. pygmaeus	Central Kalimantan	PP_535	535	F	Sanger-sequenced	This study	1. Generation of wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_536	536	F	Sanger-sequenced	This study	1. Generation of wild-born Borneo
P. pygmaeus	Central Kalimantan	PP_866	866	M	Sanger-sequenced	This study	Wild-born
P. pygmaeus	Central Kalimantan	PP_003	3	F	Sanger-sequenced	This study	Wild-born
P. pygmaeus	Central Kalimantan	PP_941	941	M	Sanger-sequenced	This study	Wild-born
P. pygmaeus	Central Kalimantan	PP_KB4204	Dolly	M	5.61	Locke *et al.* 2011	Wild-born
P. pygmaeus	Central Kalimantan	PP_KB5404	Billy	F	12.24	Locke *et al.* 2011	Wild-born
P. pygmaeus	Central Kalimantan	PP_KB5405	Dennis	M	5.61	Locke *et al.* 2011	Wild-born
P. pygmaeus	East Kalimantan	PP_496	496	F	Sanger-sequenced	This study	1. Generation of wild-born Borneo
P. pygmaeus	East Kalimantan	PP_A984	Barong	F	29.89	Chapter 4	Wild-born
P. pygmaeus	East Kalimantan	PP_A985	Panjul	M	30.13	Chapter 4	Wild-born
P. pygmaeus	North Kinabatangan	PP_A987	Tara	F	30.65	Chapter 4	Wild-born
P. pygmaeus	North Kinabatangan	PP_A988	Kala	M	31.06	Chapter 4	Wild-born
P. pygmaeus	Sarawak	PP_A939	Nonja	F	20.48	Prado-Martinez *et al.* 2013	1. Generation of wild-born Sarawak
P. pygmaeus	Sarawak	PP_A942	Gusti	F	23.12	unpubl. Prado-Martinez *et al.* 2013	1. Generation of wild-born Borneo
P. pygmaeus	Sarawak	PP_A946	Kajan	M	22.39	unpubl. Prado-Martinez *et al.* 2013	Wild-born
P. pygmaeus	Sarawak	PP_KB5406	Dinah	F	4.90	Locke *et al.* 2011	Wild-born
P. pygmaeus	South Kinabatangan	PP_5062	Ampal	M	13.81	Chapter 4	Wild-born
P. pygmaeus	South Kinabatangan	PP_A989	Micelle	F	27.30	Chapter 4	Wild-born
P. pygmaeus	West Kalimantan	PP_A983	Claus	M	29.71	Chapter 4	Wild-born

[a]mean effective whole-genome sequence coverage (base- and mapping Phred quality scores $\geqslant$ 20) (cf. Chapter 4)

Table S2. Mean MSY sequence depth compared to genome-wide coverage estimates for study males.

Species	Population	Individual	Cov_{MSY} [a]	Cov_{Gen} [b]	Source
P. abelii	Langkat	PA_A952	11.06	21.03	Prado-Martinez *et al.* 2013
P. abelii	North Aceh	PA_B018	7.93	16.31	Chapter 4
P. abelii	Batang Toru	PA_B019	6.21	16.92	Chapter 4
P. abelii	Langkat	PA_KB4661	2.79	4.76	Locke *et al.* 2011
P. abelii	Langkat	PA_KB5883	3.07	4.99	Locke *et al.* 2011
P. pygmaeus	South Kinabatangan	PP_5062	8.42	13.81	Chapter 4
P. pygmaeus	Central Kalimantan	PP_A944	12.94	23.32	Prado-Martinez *et al.* 2013
P. pygmaeus	Sarawak	PP_A946	14.07	22.39	unpubl. Prado-Martinez *et al.* 2013
P. pygmaeus	West Kalimantan	PP_A983	15.52	29.71	Chapter 4
P. pygmaeus	East Kalimantan	PP_A985	15.44	30.13	Chapter 4
P. pygmaeus	North Kinabatangan	PP_A988	16.62	31.06	Chapter 4
P. pygmaeus	Central Kalimantan	PP_KB4204	3.26	5.61	Locke *et al.* 2011
P. pygmaeus	Central Kalimantan	PP_KB5405	3.11	5.61	Locke *et al.* 2011

[a] mean effective MSY sequence coverage (base- and mapping Phred quality scores $\geqslant$ 30)

[b] mean effective whole-genome sequence coverage (base- and mapping Phred quality scores $\geqslant$ 30) (cf. Chapter 4)

Table S3. List of MSY-specific microsatellite markers. 'MSY_Region' indicates in which of the four extracted X-degenerate regions of the MSY (Table S3) the microsatellite is located. The coordinates correspond to the human Y chromosome reference sequence *GRC37*. '*StartBP_STR*' and '*EndBP_STR*' denote the start and end position of the microsatellite repeat within the '*Chromosome_Region*'. '*SeqFlankBP_Start*' and '*SeqFlankBP_End*' represent the start and end position of the extracted microsatellite sequence including 600 bp flanking region up- and downstream of the microsatellite repeat. The extracted microsatellite sequences are available from the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.[NNNN] under the '*FASTA_sequence_ID*' identifier.

MSY_Region	StartBP_STR	EndBP_STR	Repeat Motif	Type	Comments	SeqFlankBP Start	SeqFlankBP End	FASTA_sequence_ID
chrY:14170438-15795786	251282	251306	(AAAC)^6	Tetranucleotide	Forward	250682	251906	chrY:14170438-15795786:250682to251906
chrY:14170438-15795786	353528	353544	(GT)^8	Dinucleotide	Reverse: complement of AC	352928	354144	chrY:14170438-15795786:352928to354144
chrY:14170438-15795786	369631	369655	(AAAT)^6	Tetranucleotide	Forward	369031	370255	chrY:14170438-15795786:369031to370255
chrY:14170438-15795786	430399	430415	(AC)^8	Dinucleotide	Forward	429799	431015	chrY:14170438-15795786:429799to431015
chrY:14170438-15795786	451237	451255	(AG)^9	Dinucleotide	Forward	450637	451855	chrY:14170438-15795786:450637to451855
chrY:14170438-15795786	456963	456979	(AC)^8	Dinucleotide	Forward	456363	457579	chrY:14170438-15795786:456363to457579
chrY:14170438-15795786	467254	467270	(GT)^8	Dinucleotide	Reverse: complement of AC	466654	467870	chrY:14170438-15795786:466654to467870
chrY:14170438-15795786	750968	750984	(AG)^8	Dinucleotide	Forward	750368	751584	chrY:14170438-15795786:750368to751584
chrY:14170438-15795786	840873	840909	(AAT)^12	Trinucleotide	Forward	840273	841509	chrY:14170438-15795786:840273to841509
chrY:14170438-15795786	1223179	1223201	(AT)^11	Dinucleotide	Forward	1222579	1223801	chrY:14170438-15795786:1222579to1223801
chrY:14170438-15795786	1312803	1312819	(AC)^8	Dinucleotide	Forward	1312203	1313419	chrY:14170438-15795786:1312203to1313419
chrY:14170438-15795786	1337102	1337134	(AC)^16	Dinucleotide	Forward	1336502	1337734	chrY:14170438-15795786:1336502to1337734
chrY:14170438-15795786	1387994	1388034	(AGAT)^10	Tetranucleotide	Forward	1387394	1388634	chrY:14170438-15795786:1387394to1388634
chrY:14170438-15795786	1388038	1388062	(AGAT)^6	Tetranucleotide	Forward	1387438	1388662	chrY:14170438-15795786:1387438to1388662
chrY:14170438-15795786	1404285	1404309	(AT)^12	Dinucleotide	Forward	1403685	1404909	chrY:14170438-15795786:1403685to1404909
chrY:14170438-15795786	1412102	1412129	(AAT)^9	Trinucleotide	Forward	1411502	1412729	chrY:14170438-15795786:1411502to1412729
chrY:14170438-15795786	1621126	1621214	(AT)^44	Dinucleotide	Reverse: complement of AT	1620526	1621814	chrY:14170438-15795786:1620526to1621814
chrY:16470614-17686473	45312	45348	(ATTT)^9	Tetranucleotide	Reverse: complement of AAAT	44712	45948	chrY:16470614-17686473:44712to45948
chrY:16470614-17686473	93740	93788	(AGAT)^12	Tetranucleotide	Forward	93140	94388	chrY:16470614-17686473:93140to94388
chrY:16470614-17686473	141513	141529	(CT)^8	Dinucleotide	Reverse: complement of AG	140913	142129	chrY:16470614-17686473:140913to142129
chrY:16470614-17686473	141969	141989	(AC)^10	Dinucleotide	Forward	141369	142589	chrY:16470614-17686473:141369to142589
chrY:16470614-17686473	167846	167862	(GT)^8	Dinucleotide	Reverse: complement of AC	167246	168462	chrY:16470614-17686473:167246to168462

chrY:16470614-17686473	176055	176073	(GT)^9	Dinucleotide	Reverse: complement of AC	175455	176673	chrY:16470614-17686473:175455to176673
chrY:16470614-17686473	194834	194868	(GT)^17	Dinucleotide	Reverse: complement of AC	194234	195468	chrY:16470614-17686473:194234to195468
chrY:16470614-17686473	201352	201372	(GT)^10	Dinucleotide	Reverse: complement of AC	200752	201972	chrY:16470614-17686473:200752to201972
chrY:16470614-17686473	219505	219529	(ATC)^8	Trinucleotide	Forward	218905	220129	chrY:16470614-17686473:218905to220129
chrY:16470614-17686473	244400	244441	(GT)^20	Dinucleotide	Reverse: complement of AC	243800	245041	chrY:16470614-17686473:243800to245041
chrY:16470614-17686473	279128	279150	(GT)^11	Dinucleotide	Reverse: complement of AC	278528	279750	chrY:16470614-17686473:278528to279750
chrY:16470614-17686473	363042	363060	(AC)^9	Dinucleotide	Forward	362442	363660	chrY:16470614-17686473:362442to363660
chrY:16470614-17686473	418942	418962	(AG)^10	Dinucleotide	Forward	418342	419562	chrY:16470614-17686473:418342to419562
chrY:16470614-17686473	434124	434140	(GT)^8	Dinucleotide	Reverse: complement of AC	433524	434740	chrY:16470614-17686473:433524to434740
chrY:16470614-17686473	470775	470797	(GT)^11	Dinucleotide	Reverse: complement of AC	470175	471397	chrY:16470614-17686473:470175to471397
chrY:16470614-17686473	492575	492596	(GAT)^7	Trinucleotide	Reverse: complement of ATC	491975	493196	chrY:16470614-17686473:491975to493196
chrY:16470614-17686473	535822	535848	(AT)^13	Dinucleotide	Forward	535222	536448	chrY:16470614-17686473:535222to536448
chrY:16470614-17686473	539188	539206	(AC)^9	Dinucleotide	Forward	538588	539806	chrY:16470614-17686473:538588to539806
chrY:16470614-17686473	709540	709572	(AC)^16	Dinucleotide	Forward	708940	710172	chrY:16470614-17686473:708940to710172
chrY:16470614-17686473	726090	726130	(AAAG)^10	Tetranucleotide	Forward	725490	726730	chrY:16470614-17686473:725490to726730
chrY:16470614-17686473	727868	727884	(AG)^8	Dinucleotide	Forward	727268	728484	chrY:16470614-17686473:727268to728484
chrY:16470614-17686473	829314	829354	(AGAT)^10	Tetranucleotide	Forward	828714	829954	chrY:16470614-17686473:828714to829954
chrY:16470614-17686473	847116	847162	(AC)^23	Dinucleotide	Forward	846516	847762	chrY:16470614-17686473:846516to847762
chrY:16470614-17686473	849254	849322	(AAAG)^17	Tetranucleotide	Forward	848654	849922	chrY:16470614-17686473:848654to849922
chrY:16470614-17686473	891839	891857	(AC)^9	Dinucleotide	Forward	891239	892457	chrY:16470614-17686473:891239to892457
chrY:16470614-17686473	933164	933188	(AAAT)^6	Tetranucleotide	Forward	932564	933788	chrY:16470614-17686473:932564to933788
chrY:16470614-17686473	943728	943758	(ATT)^10	Trinucleotide	Reverse: complement of AAT	943128	944358	chrY:16470614-17686473:943128to944358
chrY:18837846-19267356	3575	3599	(ACAT)^6	Tetranucleotide	Forward	2975	4199	chrY:18837846-19267356:2975to4199
chrY:18837846-19267356	33806	33830	(AAAG)^6	Tetranucleotide	Forward	33206	34430	chrY:18837846-19267356:33206to34430
chrY:18837846-19267356	46217	46237	(CT)^10	Dinucleotide	Reverse: complement of AG	45617	46837	chrY:18837846-19267356:45617to46837
chrY:18837846-19267356	50169	50200	(ATCT)^7	Tetranucleotide	Reverse: complement of AGAT	49569	50800	chrY:18837846-19267356:49569to50800
chrY:18837846-19267356	62057	62075	(AC)^9	Dinucleotide	Forward	61457	62675	chrY:18837846-19267356:61457to62675
chrY:18837846-19267356	86608	86624	(AC)^8	Dinucleotide	Forward	86008	87224	chrY:18837846-19267356:86008to87224

chrY:18837846-19267356	120193	120233	(AGAT)^10	Tetranucleotide	Forward	119593	120833	chrY:18837846-19267356:119593to120833
chrY:18837846-19267356	134506	134522	(GT)^8	Dinucleotide	Reverse: complement of AC	133906	135122	chrY:18837846-19267356:133906to135122
chrY:18837846-19267356	167608	167636	(ATCT)^7	Tetranucleotide	Reverse: complement of AGAT	167008	168236	chrY:18837846-19267356:167008to168236
chrY:18837846-19267356	211628	211649	(AAT)^7	Trinucleotide	Forward	211028	212249	chrY:18837846-19267356:211028to212249
chrY:18837846-19267356	236024	236042	(AC)^9	Dinucleotide	Forward	235424	236642	chrY:18837846-19267356:235424to236642
chrY:18837846-19267356	361144	361164	(GT)^10	Dinucleotide	Reverse: complement of AC	360544	361764	chrY:18837846-19267356:360544to361764
chrY:18837846-19267356	381684	381700	(AC)^8	Dinucleotide	Forward	381084	382300	chrY:18837846-19267356:381084to382300
chrY:18837846-19267356	401214	401254	(ATCT)^10	Tetranucleotide	Reverse: complement of AGAT	400614	401854	chrY:18837846-19267356:400614to401854
chrY:21332221-21916158	33592	33614	(GT)^11	Dinucleotide	Reverse: complement of AC	32992	34214	chrY:21332221-21916158:32992to34214
chrY:21332221-21916158	37287	37305	(AC)^9	Dinucleotide	Forward	36687	37905	chrY:21332221-21916158:36687to37905
chrY:21332221-21916158	129145	129169	(AC)^12	Dinucleotide	Forward	128545	129769	chrY:21332221-21916158:128545to129769
chrY:21332221-21916158	145591	145635	(AAAG)^11	Tetranucleotide	Forward	144991	146235	chrY:21332221-21916158:144991to146235
chrY:21332221-21916158	229832	229864	(AC)^16	Dinucleotide	Forward	229232	230464	chrY:21332221-21916158:229232to230464
chrY:21332221-21916158	291300	291336	(AC)^18	Dinucleotide	Forward	290700	291936	chrY:21332221-21916158:290700to291936
chrY:21332221-21916158	310620	310636	(AG)^8	Dinucleotide	Forward	310020	311236	chrY:21332221-21916158:310020to311236
chrY:21332221-21916158	310635	310669	(GT)^17	Dinucleotide	Reverse: complement of AC	310035	311269	chrY:21332221-21916158:310035to311269
chrY:21332221-21916158	355781	355803	(AG)^11	Dinucleotide	Forward	355181	356403	chrY:21332221-21916158:355181to356403
chrY:21332221-21916158	397716	397732	(CT)^8	Dinucleotide	Reverse: complement of AG	397116	398332	chrY:21332221-21916158:397116to398332
chrY:21332221-21916158	425938	425962	(GT)^12	Dinucleotide	Reverse: complement of AC	425338	426562	chrY:21332221-21916158:425338to426562
chrY:21332221-21916158	477594	477627	(ATT)^11	Trinucleotide	Reverse: complement of AAT	476994	478227	chrY:21332221-21916158:476994to478227
chrY:21332221-21916158	514090	514106	(AC)^8	Dinucleotide	Forward	513490	514706	chrY:21332221-21916158:513490to514706

Table S4. Summary statistics for the relevant parameter estimates in BEAST for MSY.

[mya]	mean	median	ESS[a]	95% HPD[b]
TMRCA[c] all *Pongo*	0.426	0.415	3328	0.296–0.576
TMRCA all Sumatran	0.127	0.124	4197	0.087–0.174
TMRCA all Bornean	0.113	0.110	3091	0.081–0.149
ucld.mean	1.098×10^{-3}	1.092×10^{-3}	11008	7.108×10^{-4}–1.487×10^{-3}

[a]ESS: effective sample size
[b]HPD: 95% highest posterior density interval
[c]time to the most recent common ancestor

Table S5. Summary statistics for the relevant parameter estimates in BEAST for mitogenomes.

[mya]	mean	median	ESS[a]	95% HPD[b]
TMRCA[c] all *Pongo*	3.973	3.918	1198	2.352–5.572
TMRCA all Sumatran	3.973	3.918	1198	2.352–5.572
TMRCA Langkat–North Aceh/West Alas	0.969	0.950	1426	0.552–1.387
TMRCA North Aceh–West Alas	0.797	0.788	1469	0.440–1.138
TMRCA West Alas1–West Alas2	0.306	0.300	2202	0.158–0.452
TMRCA all Bornean and Batang Toru	2.405	2.401	1276	1.258–3.423
TMRCA all Bornean	0.159	0.156	1746	0.094–0.227
TMRCA South Kinabatangan–North Kinabatangan	0.040	0.040	4739	0.017–0.070
TMRCA Central/West Kalimantan–East K./Sarawak	0.146	0.144	1865	0.081–0.210
TMRCA East Kalimantan–Sarawak	0.132	0.130	1766	0.080–0.188
ucld.mean	0.011	0.010	1438	0.007–0.014

[a]ESS: effective sample size
[b]HPD: 95% highest posterior density interval
[c]time to the most recent common ancestor

Chapter 6

Whole-genome scans detect potential genetic footprints of local adaptation in orangutans (genus *Pongo*)

Maja P. Greminger[1], Alexander Nater[2,1], Javier Prado-Martinez[3], Benoit Goossens[4,5,6], Ernst Verschoor[7], Kristin Warren[8], Ian Singleton[9,10], Ivo Gut[11], Marta Gut[11], Laurentius N. Ambu[6], Carel P. van Schaik[1], Tomas Marques-Bonet[3,11], and Michael Krützen[1]

[1]Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

[2]Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

[3]CREA, Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain

[4]Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom

[5]Danau Girang Field Centre, c/o Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[6]Sabah Wildlife Department, Kota Kinabalu, Sabah, Malaysia

[7]Biomedical Primate Research Centre, Rijswijk, The Netherlands

[8]School of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, Australia

[9]Foundation for a Sustainable Ecosystem (YEL), Medan, Indonesia

[10]PanEco, Foundation f or Sustainable Development and Intercultural Exchange, Berg am Irchel, Switzerland

[11]Centro Nacional de Análisis Genómico, Barcelona, Spain

6.1 Abstract

Unravelling how organisms adapt to their natural environments is one of the major aims of evolutionary biology, but remains a challenging endeavor. Orangutans (genus: *Pongo*), the only Asian great apes, show remarkable geographic variation in various traits related to morphology, physiology, and behavioral ecology. A considerable part of this variation is likely linked to environmental differences throughout the genus' range. Here, we studied the genetic basis underlying local adaptations in orangutans by analyzing a unique dataset of whole genomes of 36 wild-born orangutans, representing the entire extant geographic distribution of both species. We followed several different strategies to detect genomic footprints of selection, including window-based genome scans to identify putative hard sweeps, the joint inference of outlier single-nucleotide polymorphisms (SNPs) and population structure using a hierarchical factor model, and the functional characterization of fixed SNPs between species. Our results, for instance, indicate that Bornean orangutans, in particular those in the northeast of the island, may exhibit adaptation pertaining to energy storage (i.e. adipose tissue) metabolism. This finding is compatible with the observed greater ability of Bornean orangutans to deposit large fat storages comparted to those on Sumatra to potentially allow physiological buffering against starvation during severe El Niño periods. We also identified several candidate genes and biological processes related to neurogenesis, which is in line with the smaller brain size of Bornean orangutans, and may also represent genetic adaptation to survive prolonged lean periods by reducing costs of metabolically expensive brain tissue. In contrast, in Sumatran orangutans, which do not face the same environmental constrains and have more favorable energy budgets, we found signatures of potential adaptive evolution within genes related to learning and adult brain plasticity, the oxytocin pathway, heart development, and hearing. We hypothesize that selective changes in some of these genes may provide Sumatran orangutans a framework for extended behavioral plasticity linked to their larger and more complex cultural repertoire and their higher sociability. Overall, our results suggest that at least some of the geographic variation in phenotypic traits of orangutans indeed represents unique genetic local adaptations. The catalogue of candidate genes and functional variants identified in this study lay a foundation for more detailed future examinations.

6.2 Introduction

Establishing the genetic basis of traits involved in local adaptation is a major goal of evolutionary biology, particularly to understand how organisms adapt to temporally and spatially varying environments (Kawecki & Ebert 2004; Stapley *et al.* 2010). Ultimately, knowledge about adaptive divergence will help to disentangle the relative importance of natural selection, genetic drift and other evolutionary forces in the process of speciation (Pardo-Diaz *et al.* 2014; Seehausen *et al.* 2014).

Until recently, studying genetic targets of selection in natural populations was methodologically limited to the examination of individual candidate genes (Sabeti *et al.* 2006). The emergence of high-throughput sequencing, however, has transformed our ability to identify the genes underpinning adaptation (Storz 2005; Akey 2009). Rapid progress in sequencing techniques, bioinformatical analysis and theoretical frameworks makes it increasingly possible to detect signatures of selection across the entire genome in population samples (Ellegren 2014). Reverse genetics (i.e. genome scan) approaches allow identifying genomic regions with footprints of selection even without a prior knowledge of the associated phenotype (e.g. reviewed in Bank *et al.* 2014; Pardo-Diaz *et al.* 2014). Extensive population-level genome scans have for instance yielded important insights into local adaptation within humans (e.g. Akey 2009; Pickrell *et al.* 2009; Pritchard *et al.* 2010; Hernandez *et al.* 2011; Grossman *et al.* 2013). However, similar examinations within taxa of non-human great apes have been scarce (but see McManus *et al.* 2015; Xue *et al.* 2015), and the required population-level genomic sequence data are just beginning to emerge (Chapters 3 and 4; Locke *et al.* 2011; Prado-Martinez *et al.* 2013; Scally *et al.* 2013; Greminger *et al.* 2014; Xue *et al.* 2015).

Orangutans (genus: *Pongo*) occur on the islands of Sumatra and Borneo that form part of the Sunda archipelago. This genus is thought to exhibit pronounced genetic local adaptations as suggested by their remarkable geographic variation in various traits related to morphology, physiology, life history, behavioral ecology, and social organization (compiled in Wich *et al.* 2009b). Because orangutans are the only non-African great apes and the phylogenetically most basal in the hominid lineage (Groves 2001), the genus *Pongo* is of high interest for the understanding of the adaptive evolutionary history of great apes in general (Locke *et al.* 2011; Prado-Martinez *et al.* 2013). The well-documented variation in orangutan phenotypic traits (van Schaik *et al.* 2009b; Wich *et al.* 2009b; there is insufficient data on one Bornean subspecies, *Pongo P. pygmaeus*) largely follows a west–east gradient (Figure 1) across the genus' entire range from northern Sumatra (*Pongo abelii*) via western and central Borneo (*Pongo pygmaeus wurmbii*) to eastern and northern Borneo (*Pongo pygmaeus morio*). Most of this phenotypic variation is almost certainly at least partially related to habitat differences along the same west–east gradient (Krützen *et al.* 2011; Wich *et al.* 2011b), in particular to the temporal and spatial stability of food supply, and perhaps also the abundance of large predators such as tigers (van Schaik *et al.* 2009b).

Rainforests on northern Sumatra are generally better habitat for orangutans than Bornean forests for several reasons. The overall forest productivity is for instance higher on northern Sumatra, most likely due to the younger, by volcanism enriched soils and the lower levels of cloudiness (Husson *et al.* 2009; Marshall *et al.* 2009; Wich *et al.* 2011b). Furthermore, fruit abundance is temporally more stable in northern Sumatra, while orangutans in northeastern Borneo have to cope with marked fluctuations in fruit availability, including mast fruiting events where short periods of overabundance of fruit are followed by extended periods of low fruit production (Wich *et al.* 2006; Morrogh-Bernard *et al.* 2009; Kanamori *et al.* 2010; Wich *et al.* 2011b). Especially in east and northern Borneo (range of *P. p. morio*) habitat quality is further lowered by effects caused by the El Niño-Southern Oscillation phenomenon (ENSO) (Philander 1983), which likely has been active since at least the Late Pleistocene (Philander 1983; Allan *et al.* 1996; Nipperess 2015). Prolonged droughts and forest fires that recur at 2–10 year intervals in conjunction with ENSO events (Dilley & Heyman 1995) pose major survival threats to the orangutans in this region (MacKinnon *et al.* 1996; Knott 1998; Delgado & van Schaik 2000). The unpredictable ENSO periods lead to very low abundance of edible food and thus periods of extreme food scarcity, during which orangutans are forced to rely on low-energy (fallback) foods such as inner bark, leaves, and other vegetation to a great extent (Knott 1998; Morrogh-Bernard *et al.* 2009). Taken together, this environmental gradient likely causes considerable adaptive pressure on orangutans from different regions.

Linked to the differences in habitat quality, orangutan population densities decrease from west to east (Husson *et al.* 2009; Marshall *et al.* 2009). Furthermore, in the same direction, we see an increase in mandibular robusticity and probably tooth enamel thickness (Taylor 2006; Taylor 2009), whereas both absolute and relative brain size of female orangutans significantly decrease (Taylor & van Schaik 2007; C. P. van Schaik 2010, unpublished data). Orangutans also differ in their physiology. Studies measuring ketone bodies excreted in urine of wild individuals indicate a greater tendency of Bornean orangutans to deposit large fat storages (Knott 1998; Wich *et al.* 2006). This finding is supported by observations in captivity where obesity in response to long-term food overabundance is more commonly seen among Bornean orangutans, though there has not yet been any formal study of this (Dierenfeld 1997; van Schaik *et al.* 2009b). Significant variation is also found in life history. It appears that Bornean orangutans exhibit a faster-paced life history than their Sumatran counterparts with shorter interbirth intervals (Wich *et al.* 2009a), earlier reduced association with the mother and younger age at first birth (van Noordwijk *et al.* 2009; C. P. van Schaik 2010, unpublished data). These differences are again most pronounced between Sumatran (*P. abelii*) and the northeastern Bornean orangutans (*P. p. morio*). Among the most striking examples of the broad variation in orangutan phenotypes are also the higher sociability and social tolerance of Sumatran orangutans (van Schaik 1999; van Schaik 2004; Knott *et al.* 2008; Mitra Setia *et al.* 2009; Weingrill *et al.* 2011), as well as their larger cultural repertoire comparted to Bornean orangutans (van Schaik 2004; van Schaik *et al.* 2009a; Krützen *et al.* 2011).

The remarkable variation in orangutan phenotypic traits offers a great opportunity to study the interaction between environmental pressures and genetic adaptation both between and within two closely-related great ape species. So far, signatures of positive selection in orangutan genomes have mainly been examined at the genus-wide level, i.e. selection that has acted along the *Pongo* lineage since their divergence from the last common ancestor with the African great apes (Locke *et al.* 2011; Ma *et al.* 2013). Here, we present the first genome-wide scans for positive selection within the genus *Pongo*, relying on a unique dataset of whole genomes of 36 wild-born orangutans representing the entire geographic distribution of both species (Figure 1, Table 1). We used several complementary approaches to identify potential footprints of local adaptation in orangutan genomes and investigate what differentiates the two species at the level of the genome. We analyzed patterns of genetic variation among orangutans to identify genomic regions with signatures of strong hard sweeps, i.e. genetic hitchhiking associated with selection driving a single newly arising or very rare beneficial mutation rapidly to high frequency (Maynard Smith & Haigh 1974). Furthermore, we functionally characterized single SNPs that were fixed between species, as well as applied a hierarchical factor model (PCAdapt; Duforet-Frebourg *et al.* 2015) to jointly infer outlier SNPs and population structure.

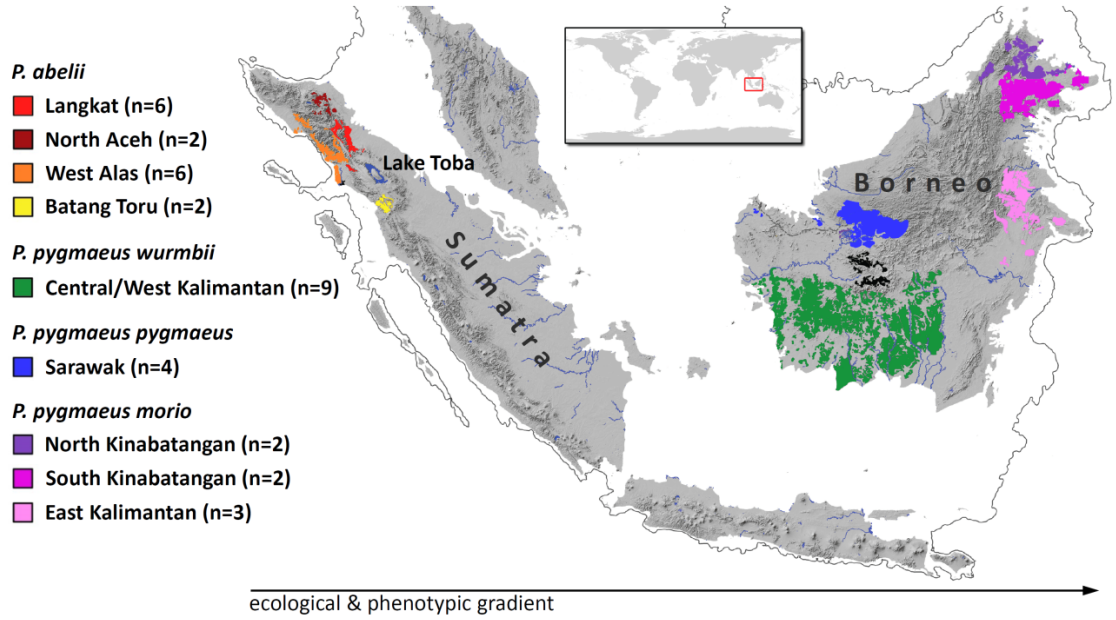


Figure 1. Map of current orangutan distribution in Sundaland. Extant orangutan populations are highlighted by different colors. Whole-genome sample sizes are given in parentheses. The thin grey line indicates the extent of the exposed Sunda shelf during the last glacial maximum (19–26 ka, ~120 meters below current sea level).

6.3 Materials and Methods

Sampling and whole-genome sequencing

Our dataset encompassed whole genomes of 16 Sumatran and 20 Bornean orangutans with known population provenance (Table 1, Figure 1). Details on study individuals, whole-genome sequencing, and bioinformatical procedures can be found in Chapter 4. In brief, to obtain a complete representation of extant orangutan populations (Figure 1), we complemented previous sequencing efforts by sequencing genomes of 11 wild-born orangutans. We analyzed the novel genomes together with 20 genomes previously sequenced by Locke *et al.* (2011, n=10) and Prado-Martinez *et al.* (2013, n=10). Furthermore, we added five unpublished genomes previously sequenced by Prado-Martinez *et al.* (2013). Mean effective sequencing depth, estimated from filtered BAM files, varied among individuals from 4.8–12.2x (Locke *et al.* 2011) to 13.7–31.1x (Chapter 4, Prado-Martinez *et al.* 2013). We produced high quality genotypes for all individuals for each position in the genome, applying the same filtering criteria for SNP and non-polymorphic positions (Chapter 4). Among all individuals, we identified 30,640,634 SNPs.

Table 1. Overview sampling for whole-genome sequencing.

Species	Population	This study (Chapter 4) [a]	Prado-Martinez *et al.* 2013	Locke *et al.* 2011	Total
P. abelii	Langkat	0	4	2	6
P. abelii	North Aceh	1	1	0	2
P. abelii	West Alas	4	0	2	6
P. abelii	Batang Toru	1	0	1	2
P. p. morio	South Kinabatangan	2	0	0	2
P. p. morio	North Kinabatangan	2	0	0	2
P. p. morio	East Kalimantan	2	0	1	3
P. P. pygmaeus	Sarawak	2	1	1	4
P. p. wurmbii	Central/West Kalimantan	2	4	3	9

[a] including unpublished genomes sequenced by Prado-Martinez *et al.* (2013)

Identification of candidate loci for positive selection

To study the genetic basis of phenotypic differences among orangutans, we applied both single SNP and window-based approaches, which are outlined below. While the analysis of single SNPs may provide important insights into actual functional changes in coding sequences, window-based genome scans allow detecting selective sweep patterns and can account to a certain degree for demographic factors and other sources of variation by averaging statistics over a larger number of SNPs (Excoffier *et al.* 2009; Bazin *et al.* 2010; Lawson *et al.* 2012).

Window-based whole-genome scans

We first searched for species-specific hard sweep patterns through windowed whole-genome scans of several population genetic summary statistics. For this analysis, we excluded the Batang Toru population, which is the only remaining population south of Lake Toba on Sumatra. Batang Toru would need to be treated as a separate unit since they are highly distinct from the other Sumatrans for both the autosomal and mitochondrial genome (Chapters 4 and 5). Unfortunately, the low sample size (n = 2) for Batang Toru does currently not permit such an analysis.

We performed genome scans on a per-chromosome basis in sliding windows with 100 kb length and 25 kb step size using custom Perl scripts (available upon request), For each genome position, we required at least 10 genotypes (i.e. individuals) per species with a minimal sequence depth of 5x, otherwise sites were coded as missing. Windows with less than 33.3 kb (1/3) valid sites after this filtering were excluded from analysis. For each window, we calculated the following summary statistics as mean values over all available sites within the window: (i) between-species population differentiation (F_{ST}), (ii) the proportion of fixed differences between species (d_f), (iii) mean pairwise sequence divergence between species (d_{xy}), (iv) within-species nucleotide diversity (π) based on the mean number of pairwise sequence differences, (v) Watterson estimator of θ within-species based on the proportion of segregating sites (θ_W) (Watterson 1975), and (vi) within-species Tajima's d (Tajima 1989).

F_{ST} is a relative measure of differentiation dependent on the between- and within-species genetic diversity, and was estimated according to Nei (1973). d_f was calculated as the number of fixed between-species differences within each window divided by the total number of available sequence for this window. A site was considered to be fixed between species if all Bornean orangutans were homozygous for one allele and all Sumatran orangutans for the other allele. d_{xy} was also measured per site within each window, as mean pairwise nucleotide difference between all pairs of chromosomes from different species. Within-species π was calculated as the mean number of pairwise differences per site between all pairs of chromosomes within a species.

We identified putative species-specific hard sweeps based on the following rationale: as the frequency of a beneficial allele increases in one species, linked genetic variation decreases by genetic hitchhiking and species differentiation increases (Nielsen *et al.* 2005b; Pritchard *et al.* 2010; Olson-Manning *et al.* 2012). A species-specific selective sweep region should therefore be characterized by highly elevated levels of F_{ST} between Bornean and Sumatran orangutans and strongly reduced π in one species but not in the other (Lewontin & Krakauer 1973). We chose to take windows with an average F_{ST} above the 99th percentile of the empirical distribution over the whole genome as candidates for putative selective sweep regions (n = 925). For each of these windows, we checked if π was significantly reduced in Bornean and/or Sumatran orangutans by deriving empirical significance thresholds for each species based on the π values of the 0.1%, 1% and 5% left tail of the π distribution over all 92,547 windows of

the genome scan (Table 2, Online Supporting Table SO1). For each species, we considered a window to be a putative sweep region if its F_{ST} estimate was among the highest 1% and if π was significantly ($P < 0.05$) reduced in that species but not in the other. We ranked windows according to highest common logarithm of the absolute between-species π ratio $[Log_{10}(\pi_{PP}/\pi_{PA})]$. For illustration of the genome scans, we plotted windowed summary statistics for each chromosome in R using the package 'ggplot2' (Wickham 2009).

Table 2. π-values for three significance levels of within-species π reduction.

	***P*-value = 0.01**	***P*-value = 0.10**	***P*-value = 0.05**
P. pygmaeus	0.000152757	0.000338559	0.000601208
P. abelii	0.000407843	0.000712784	0.001060288

Candidate gene information

We identified protein-coding genes located within putative sweep regions with the BioMart web-interface (Kasprzyk 2011) of the Ensembl genome browser (http://www.ensembl.org/biomart/), searching the '*Pongo abelii* genes' dataset (Ensembl release 78). We further gathered detailed information on identified protein-coding genes using GeneALaCart (LifeMap Sciences, Inc.), which allows extracting information on a large number of genes from the GeneCards encyclopedia—an integrated database of information dealing with human genes (https://genealacart.genecards.org/; last accessed March 5th 2015; Safran *et al.* 2010). We obtained GeneCards summaries of gene function and disease association annotations from the following major knowledge databases: Entrez Gene of the National Center for Biotechnology Information (NCBI), UniProt Knowledgebase (UniProtKB/Swiss-Prot), The Human Malady Compendium (MalaCards), and DISEASES database (disease-gene associations mined from literature).

Analyses at SNP-level

Fixed SNPs between species

We investigated fixed SNPs between species in more detail, i.e. SNPs for which all Borneans were homozygous for one allele and all Sumatrans for the other (again excluding the Batang Toru population). We only considered SNP positions which were covered by at least 10 genotypes per species with a minimal sequence depth of 5x (n = 27,037,765 SNPs) and identified fixed SNPs with custom R scripts. To characterize the effects of fixed SNP variants on genes, transcripts, and protein sequence we used the Variant Effect Predictor (McLaren *et al.* 2010) implemented in the Ensembl genome browser (http://www.ensembl.org/info/docs/tools/vep/). For all genes containing at least one non-synonymous fixed SNP, we obtained again detailed functional information with GeneALaCart.

PCAdapt analysis

Finally, we identified candidate SNPs potentially involved in local adaptation by jointly inferring population structure and outlier loci with the program PCAdapt v1.6 (Duforet-Frebourg *et al.* 2014). We run the *fast* implementation of PCAdapt that is based on principal component analysis (PCA) and suitable for genome-wide datasets (Duforet-Frebourg *et al.* 2015). PCAdapt captures population structure by K principal components (PCs) and identifies outlier loci as SNPs that are excessively related to one of the PCs.

We performed two separate PCAdapt analyses, one between orangutan species (including Batang Toru) and one within Bornean orangutans, as we were particularly interested in local adaptations of *P. p. morio*. We produced input files with custom-made perl and bash scripts, requiring a minimal sequence depth of 5x per genotype and excluding SNPs with more than one missing genotype per species. This left us with 15,985,632 SNPs for the between-species analysis and 10,775,794 SNPs for the analysis within Bornean orangutans (polymorphic SNPs only). In both analyses, we ranked SNPs based on their squared loadings (p^2) (Duforet-Frebourg *et al.* 2015) with the first PC, i.e. the correlations between the genotypes at a given SNP and PC1. We ranked SNPs according to PC1 because we were mainly interested in selection occurring along this axis, which separated Bornean and Sumatran orangutans in the between-species analysis, and *P. p. wurmbii* from *P. P. pygmaeus* and *P. p. morio* in the within-Borneo analysis, respectively (see Results). For the between-species analysis, we selected the 1% highest-ranking SNPs as outlier candidates, for the analysis within Bornean orangutans we used only the top 0.5% SNPs since considerably fewer selection targets are expected within than between species.

Gene ontology enrichment analyses

To examine whether genes within putative selective sweep regions (i.e. candidate genes) were enriched for any particular biological process, we performed an analysis of Gene Ontology (GO) terms using the R package 'gProfileR' of the g:Profiler toolkit (Reimand *et al.* 2007; Reimand *et al.* 2011). Significance was assessed by comparing the candidate genes with a background list of all possible genes, i.e. all protein-coding genes (n = 12,866) located within any window of the genome scan. We applied the default g:SCS method (Reimand *et al.* 2011) for computing multiple testing correction for *P*-values gained from GO enrichment analysis.

We also performed GO enrichment analyses for the identified candidate SNPs of the two PCAdapt analyses as well as for all fixed SNPs between species. For these SNP-based GO analyses, we used the program GOWINDA (Kofler & Schlötterer 2012)—a software that was designed for genome-wide association studies. Classical GO analyses may be biased as longer genes typically have more SNPs, thus a higher probability of being sampled (Kofler & Schlötterer 2012). Permutation tests implemented in GOWINDA take this into account and allow for an unbiased analysis of gene set enrichment (Kofler & Schlötterer 2012). To obtain

an orangutan-specific gene set file of GO terms, we downloaded biological process GO terms for all protein-coding genes in the orangutan genome from the BioMart web-interface (accessed October 5th 2014) and converted the file to the format required by GOWINDA in R. We ran GOWINDA for each candidate SNP set separately as following: the list of candidate genes was built from genes which contained at least one candidate/fixed SNP within a window of 5,000 bp upstream and downstream of the gene. Including these flanking regions ensured to capture also SNPs within close-by regulatory elements (Blanchette *et al.* 2006; ENCODE Project Consortium 2012). The background list of genes for significance assessment was derived from all SNPs used in the PCAdapt analysis or to identify the fixed SNPs, respectively. We applied the recommended more conservative '–gene flag', which assumes that all SNPs within a gene are completely linked. Significance thresholds ($P < 0.05$) after false-discovery rate (FDR) correction were obtained empirically based on 100,000 simulations.

6.4 Results

Window-based genome scans

The window-based scans revealed a heterogeneous genomic landscape of species differentiation between Bornean and Sumatran orangutans, with regions of highly elevated F_{ST} and d_f along all chromosomes (Figure 2, Supporting Figure S1, Online Supporting Table SO1). In the 925 windows above the 99th percentile of the empirical F_{ST} distribution (Online Supporting Table SO2), F_{ST} ranged from 0.491–0.788, while the genomic mean F_{ST} between the two orangutan species was 0.201 (Table 3). Strikingly, in 376 (40.65%) of these top 1% F_{ST} windows, genetic diversity was significantly ($P < 0.05$) reduced in both orangutan species, indicating considerable genus-wide background selection due to reduced local recombination rate in both species, or parallel sweeps (see Discussion).

Table 3. Mean values of population genomic summary statistics[a].

Statistic	Total	*P. pygmaeus*	*P. abelii*[b]
F_{ST}	0.201 ± 0.084	/	/
d_f (x10^{-3})	0.030 ± 0.110	/	/
d_{xy} (x10^{-3})	3.088 ± 1.083	/	/
π ± s.d. (x10^{-3})	2.505 ± 0.939	1.741 ± 0.924	2.286 ± 0.886
θ_W[b] ± s.d. (x10^{-3})	2.125 ± 0.715	1.231 ± 0.620	1.759 ± 0.698
Tajima's d (x10^{-3})	0.380 ± 0.336	0.509 ± 0.364	0.527 ± 0.291

[a] ± standard deviation

[b] excluding the Batang Toru population

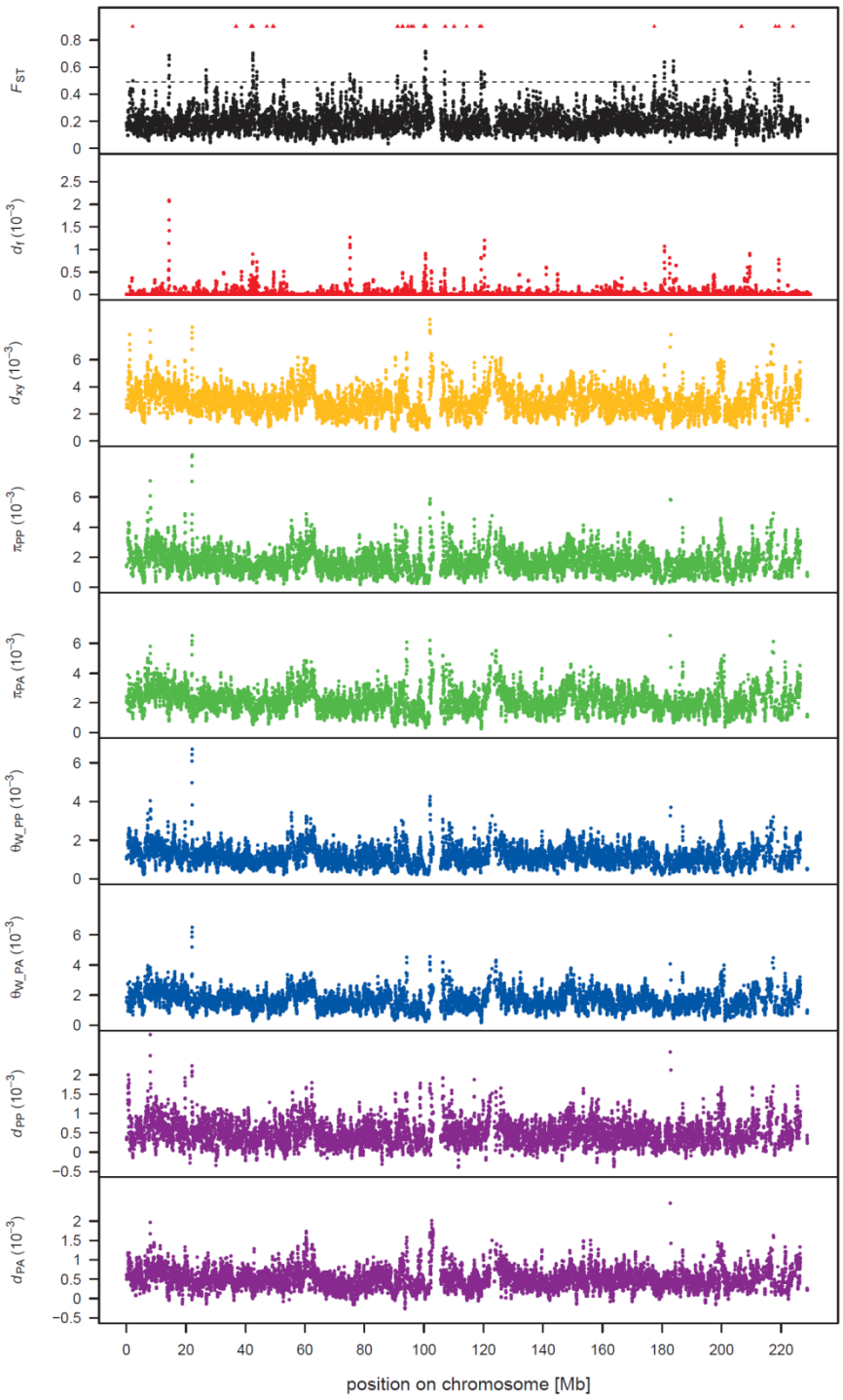


Figure 2. Distribution of windowed population genetic summary statistics along chromosome 1 (example chromosome). Summary statistics (y-axis) were averaged in windows of 100 kb length, sliding in 25 kb steps along the chromosome (x-axis). Plotted are the between-species population differentiation (F_{ST}), the density of fixed differences between species per base pair (d_f), the mean pairwise between-species sequence divergence (d_{xy}), the within-species nucleotide diversity for Bornean (π_{PP}) and Sumatran orangutans (π_{PA}), the within-species Watterson estimator ($\theta_{W\_PP/PA}$), and within-species Tajima's d ($d_{PA/PP}$). The dashed black line indicates windows above the 99th percentile of the empirical F_{ST} distribution. The small red triangles at the top of the F_{ST} plot denote chromosome positions of fixed non-synonymous SNPs between species.

Candidate genes in Bornean orangutans

In 161 of the top 1% F_{ST} windows (n = 925), π was significantly ($P < 0.05$) reduced exclusively in Bornean orangutans, designating putative species-specific selective sweeps (Figure 3, Online Supporting Table SO3). Windows clustered in 71 genomic regions with an average length of 132 kb (median = 125 kb; range = 100 to 600 kb). In total, candidate sweep regions covered 9.5 Mb or 0.33% of the haploid autosomal genome. Within putative selective sweep regions, we identified 113 protein-coding genes of which 15 were uncharacterized novel genes in the genus *Pongo* (identified by Ensembl Gene Build). We did not find significant ($P < 0.05$) GO enrichment categories for protein-coding genes located within putative selective sweep regions after correction for multiple testing, which might be explained by two main factors: first, sweep regions also contain genes that are likely not target of selection leading to dilution of the signal. Second, out of 20,424 protein-coding genes in the orangutan genome only 13,229 genes have an assigned 'biological processes' GO term. Detailed information on all candidate genes, including summaries of their potential function and disease associations in humans and other animals, are provided in Online Supporting Table SO4.

Brain development

Candidate genes related to local adaptation in Bornean orangutans include genes associated with brain development and structure. *FOXG1* (previously known as brain factor 1) encodes a transcription repression factor that plays an important role in the development of the telencephalon as well as in the establishment of regional subdivision of the developing brain by controlling neurogenesis (Hébert & McConnell 2000; Martynoga *et al.* 2005; Kortüm *et al.* 2011). Mutations in *FOXG1* cause severe microcephaly (Kortüm *et al.* 2011). Another candidate gene is *POMGNT1*, which encodes a type II transmembrane protein. Missense mutations in *POMGNT1* cause muscle-eye-brain disease and Walker–Warburg syndrome characterized by severe brain and eye abnormalities (Hanemaaijer *et al.* 2009; Saredi *et al.* 2012).

Lipid and glucose metabolism

We also identified several candidate genes associated with lipid and glucose metabolism. For example, PIK3R3 directly interacts with the Insulin-like growth factor 1 receptor in humans, RABL3 has been associated with obesity (Wilton & Matthews 1996; Almon *et al.* 2009; Comuzzie *et al.* 2012), and SCAP encodes an escort protein required for cholesterol as well as lipid homeostasis (The UniProt Consortium 2015). Furthermore, two out of the nine protein-coding genes located within the 20 highest-ranking putative selective sweep windows (Tables 4 and 5) are related to lipid and glucose metabolism. SLC6A2, a neurotransmitter transporter, regulates norepinephrine homeostasis. Norepinephrine is a stress hormone and directly increases heart rate, triggering the release of glucose from energy stores, and increasing blood flow to skeletal muscle. Increase in norepinephrine levels may be an initial signal for metabolic changes in early starvation (e.g. Zauner *et al.* 2000; Patel *et al.* 2002; Goldstein *et*

al. 2011; Gagnon & Anini 2013). Mutations in the SLC6A2 gene also cause orthostatic intolerance, a syndrome characterized by lightheadedness, fatigue, altered mentation and syncope. Another gene located within the 20 top-ranking sweep windows was KIAA1109. The function of KIAA1109 is largely unknown, but it has been associated with type 1 diabetes (The Wellcome Trust Case Control Consortium 2007; Barrett *et al.* 2009), as well as with susceptibility to celiac disease in humans (van Heel *et al.* 2007; Zhernakova *et al.* 2007). The encoded protein of KIAA1109 is similar to a hamster protein that is known to play a role in adipocyte differentiation (Wei *et al.* 2006).

Candidate genes in Sumatran orangutans

For Sumatran orangutans, we detected 81 putative selective sweep regions (212 windows; Online Supporting Table SO5). Regions had an average length of 163 kb (median = 125 kb; range = 100 kb to 1.25 Mb) and covered 13.2 Mb or 0.46% of the autosomal genome (Figure 3). A total of 163 protein-coding genes, of which 5 were uncharacterized novel genes in the genus *Pongo*, were located within putative sweep regions. As for Bornean orangutans, we found no significantly ($P < 0.05$) enriched GO categories for candidate genes. Detailed functional information on all candidate genes for Sumatran orangutans is given in Online Supporting Table SO6.

Heart function

Two out of five genes located within the 20 highest-ranking putative sweep windows are related to heart function (Tables 4 and 5). One of them, *EPHA3*, is located within the largest putative selective sweep region in Sumatran orangutans, covering 17 out of the 20 top windows and spanning in total 1.25 Mb on chromosome 3 (chr3: 56,150,000 – 57,400,000; Online Supporting Table SO5). *EPHA3* is the only protein-coding gene located within this region and encodes a *protein*-tyrosine kinase belonging to the ephrin receptor subfamily. *EPHA3* receptor signaling plays a critical role in heart development, in particular in the formation of the atrioventricular canal and septum (Stephen *et al.* 2007; Frieden *et al.* 2010). The second gene related to heart function in the top-ranking windows was the *transmembrane protein 43 (TMEM43)*. Defects in *TMEM43* cause arrhythmogenic right ventricular cardiomyopathy type 5, which is characterized by ventricular tachycardia, heart failure, sudden cardiac death, and fibrofatty replacement of cardiomyocytes (Merner *et al.* 2008; Baskin *et al.* 2013).

Social behavior, learning and lactation

We also identified candidate genes related to social behavior, lactation, and learning. Among the genes in the 20 top-ranking sweep windows (Tables 4 and 5) was for example *TNPO1*, which is a carrier protein transporting substrates between the cytoplasm and the nucleus. *TNPO1* is the primary carrier for oxytocin receptor (*OXTR*) nuclear transport (Di Benedetto *et al.* 2014), which responds to the neuropeptide oxytocin to stimulate for example lactation and social behavior (Kosfeld *et al.* 2005; Heinrichs *et al.* 2009). Furthermore, oxytocin shapes

the physical development of the human neocortex as well as social learning (Leuner *et al.* 2012; reviewed in Carter 2014).

Other candidate genes in Sumatran orangutans are of critical importance for learning and adult brain plasticity. For example, the gene *ATAD1* regulates the surface expression of *AMPA* receptors, thereby playing a central role in hippocampus-dependent learning and memory (Whitlock *et al.* 2006; Keifer & Zheng 2010; Mitsushima *et al.* 2011). Another gene, *T-brain-1* (*TBR1*), is a crucial neuron-specific transcription factor for forebrain development. The up-regulation of *TBR1* expression requires in turn the activation of *AMPA* receptors, regulated by the candidate gene *ATAD1*. *TBR1* seems to play an important role in adult mouse brain in response to neuronal activation to modulate gene transcription required for neural plasticity (Chuang *et al.* 2014; Huang *et al.* 2014). In humans, *TBR1* was found to be associated with educational attainment in a genome-wide association study (Rietveld *et al.* 2014). Among the Sumatran candidate genes was also the *cell division protein kinase 6* (*CDK6*), which is a key molecular regulator of neurogenesis in the adult brain, and thus of brain plasticity (Beukelaers *et al.* 2011; Caron *et al.* 2014). *CDK6* is essential for the production of neurons within the hippocampus. Due to its activities, it also delays senescence (Ruas *et al.* 2007).

Figure 3. Candidate selective sweep regions across the genome. For each autosomal chromosome (1–22), the distribution of windowed FST is plotted (grey dots). The x-axis represents the genomic position, the y-axis the average FST values for 100-kb windows. Putative selective sweep windows are colored in blue and pink for Sumatran and Bornean orangutans, respectively. These windows rank among the top 1% FST windows (dashed black line) and show significantly ($P < 0.05$) reduced genetic diversity exclusively in the respective species. Top-ranking FST windows for which genetic diversity was significantly reduced in both species are colored in yellow. Small black triangles denote chromosome locations of fixed non-synonymous SNPs.

Table 4. List of the 20 highest-ranking windows of putative selective sweeps in Bornean (*P. pygmaeus*) and Sumatran (*P. abelii*) orangutans. Four putative sweep regions in Bornean orangutans lacked annotated protein-coding genes, suggesting that unknown functional elements might be target of selection.

Chromo	Start	End	F_{ST}	d_{xy}	π_{PP}	π_{PA}	*P*-value[a]	$Log_{10}(\pi_{PP}/\pi_{PA})$	Gene(s)
P. pygmaeus									
chr1	52825000	52925000	0.50450	0.00253	0.00023	0.00132	**	-0.76588	*DENND1B*
chr2b	13100000	13200000	0.49600	0.00260	0.00026	0.00137	**	-0.72085	*CLASP1*
chr2b	89725000	89825000	0.55789	0.00246	0.00022	0.00109	**	-0.69826	/
chr3	34000000	34100000	0.50791	0.00358	0.00022	0.00194	**	-0.94471	
chr3	34025000	34125000	0.58823	0.00414	0.00027	0.00175	**	-0.80897	/
chr3	34050000	34150000	0.65083	0.00413	0.00026	0.00141	**	-0.73615	
chr3	34075000	34175000	0.70944	0.00389	0.00021	0.00106	**	-0.70983	
chr4	93025000	93125000	0.59039	0.00335	0.00028	0.00136	**	-0.69171	*TIGD2*
chr4	126800000	126900000	0.50686	0.00228	0.00015	0.00123	***	-0.92112	*KIAA1109*
chr4	126825000	126925000	0.52654	0.00211	0.00013	0.00108	***	-0.91893	
chr4	148700000	148800000	0.51863	0.00315	0.00031	0.00156	**	-0.70553	ENSPPYG00000015089
chr4	158450000	158550000	0.55293	0.00263	0.00021	0.00121	**	-0.75850	*TIGD4*
chr8	36200000	36300000	0.51582	0.00601	0.00042	0.00314	*	-0.87749	/
chr8	36225000	36325000	0.55842	0.00546	0.00034	0.00255	**	-0.88088	
chr12	96375000	96475000	0.57643	0.00302	0.00021	0.00132	**	-0.79573	*VEZT*
chr12	96400000	96500000	0.58628	0.00305	0.00021	0.00129	**	-0.79463	
chr13	92425000	92525000	0.50996	0.00457	0.00022	0.00251	**	-1.06146	/
chr16	42725000	42825000	0.51177	0.00343	0.00011	0.00191	***	-1.23547	
chr16	42750000	42850000	0.52615	0.00380	0.00014	0.00202	***	-1.14615	*LPCAT2/ SLC6A2*
chr16	42775000	42875000	0.52507	0.00374	0.00026	0.00190	**	-0.87111	
P. abelii									
chr3	1525000	1625000	0.53167	0.00358	0.00190	0.00041	**	0.66860	*TMEM43/ CHCHD4/XPC*
chr3	1550000	1650000	0.50944	0.00300	0.00173	0.00035	***	0.69577	
chr3	56150000	56250000	0.50361	0.00562	0.00341	0.00056	**	0.78090	
chr3	56175000	56275000	0.56290	0.00579	0.00311	0.00034	***	0.96707	
chr3	56200000	56300000	0.56483	0.00651	0.00341	0.00043	**	0.89936	
chr3	56225000	56325000	0.52815	0.00609	0.00341	0.00058	**	0.76594	*EPHA3*
chr3	56675000	56775000	0.56531	0.00477	0.00236	0.00043	**	0.73830	
chr3	56800000	56900000	0.52658	0.00489	0.00300	0.00027	***	1.04298	
chr3	56825000	56925000	0.51931	0.00472	0.00295	0.00027	***	1.03172	
chr3	56850000	56950000	0.51605	0.00466	0.00296	0.00025	***	1.06841	
chr3	56900000	57000000	0.54181	0.00563	0.00297	0.00057	**	0.71896	
chr3	56925000	57025000	0.54631	0.00580	0.00304	0.00056	**	0.73193	
chr3	56950000	57050000	0.55603	0.00604	0.00306	0.00058	**	0.71902	
chr3	56975000	57075000	0.57688	0.00601	0.00301	0.00041	***	0.87137	
chr3	57000000	57100000	0.50828	0.00553	0.00364	0.00027	***	1.12386	
chr3	57125000	57225000	0.50876	0.00477	0.00302	0.00033	***	0.95592	
chr3	57150000	57250000	0.53021	0.00490	0.00286	0.00036	***	0.90208	
chr3	57175000	57275000	0.52962	0.00450	0.00258	0.00037	***	0.84250	
chr3	57225000	57325000	0.49330	0.00398	0.00245	0.00044	**	0.74114	
chr5	72875000	72975000	0.49441	0.00193	0.00121	0.00019	***	0.80272	*TNPO1*

[a]Significance level of within-species π reduction (*: $P < 0.5$; **: $P < 0.1$; ***: $P < 0.01$)

Table 5. Function of the genes in Table 4. A more detailed description of the genes is provided in the Online Supporting Tables SO4 and S06.

Gene	Potential function
P. pygmaeus	
DENND1B	activates *Rab35* which is a key regulator of intracellular membrane trafficking and may indirectly regulate neurite outgrowth
CLASP1	involved in the regulation of microtubule dynamics at the kinetochore and throughout the spindle
TIGD2	unknown
KIAA1109	potential role in adipocyte differentiation; associated with susceptibility to celiac disease
ENSPPYG00000015089	unknown; uncharacterized novel protein-coding gene
TIGD4	unknown
VEZT	transmembrane protein; pivotal functions in the establishment of adherens junctions; role in gastric cancer
SLC6A2	neurotransmitter transporter regulating norepinephrine homeostasis
LPCAT2	potential function in membrane biogenesis and production of the platelet-activating factor in inflammatory cells
P. abelii	
TMEM43	transmembrane protein; defects cause arrhythmogenic right ventricular cardiomyopathy type 5
CHCHD4	functions as chaperone and catalyzes the formation of disulfide bonds in substrate proteins
XPC	component of the nucleotide excision repair pathway; recognizes a wide spectrum of damaged DNA
EPHA3	receptor tyrosine kinase playing a critical role in heart development
TNPO1	primary carrier protein for oxytocin receptor to stimulate lactation and social behavior

Differentiation of Bornean and Sumatran orangutans at SNP-level

In a complementary approach to the window-based scans to identify putative selective sweeps, we characterized what differentiates the two orangutan species at the single SNP level, either by genetic drift or the impact of directional selection.

Fixed SNPs between orangutan species

Out of 27,037,765 analyzed autosomal SNPs, 123,023 SNPs (0.455%) were completely fixed for different alleles in Bornean and Sumatran orangutans (Online Supporting Table SO7). Because we lack the ancestral state information for SNPs, we cannot make any statements about in which orangutan species the derived allele got fixed. Of all fixed SNPs, 39.9% were located within 5 kb of a protein-coding gene, indicating an enrichment of fixed SNPs in gene and regulatory regions (Online Supporting Table SO8). Gene ontology analysis of protein-coding genes containing fixed SNPs (3,889 genes) revealed statistically significant (P FDR < 0.05) enrichment of 19 biological GO categories (Online Supporting Table SO10). In Table 6, we list the GO categories which were also significantly enriched in our between-species PCAdapt analysis (see below). Thirteen GO terms were significant in both analyses; another two terms were almost significant in the GO analysis of PCAdapt outliers, scoring within the

50 top-ranking terms (out of 6,409 terms). The significantly enriched GO terms were associated with brain development (n = 2), skeletal development (n = 3), metabolism (n = 5), organismal development (n = 4), and regulation of transcription (n = 3). Enriched gene ontologies include for example two terms possibly associated with differences in diet between Bornean and Sumatran orangutans, i.e. the sensory perception of taste and the response to stilbenoid. Stilbenoids are a class of plant phenolics occurring in the wood and fruits of several plant families, including tropical Dipterocarpaceae and Gnetaceae (Sotheeswaran & Pasupathy 1993). Orangutan fallback food is higher in phenolics than fruit pulp and seeds (Leighton 1993).

Table 6. Significantly enriched gene ontology (GO) terms between Sumatran and Bornean orangutans. Listed GO terms were significantly enriched in the analysis of both fixed SNPs (Online Supporting Table SO9) and the 1% top-hit SNPs of the between-species PCAdapt analysis (Online Supporting Table SO12), respectively. We report only GO terms that are related to biological processes.

GO term	GO description	*P* FDR[a] $_{fixedSNP}$	*P* FDR[a] $_{PCAdapt}$	No. of genes[b] $_{fixedSNP}$	No. of genes[b] $_{PCAdapt}$	Rank[c]
Brain development						
GO:0021797	forebrain anterior/posterior pattern specification	0.00165	0.08852[§]	5/5	3/5	42
GO:0021938	smoothened signaling pathway involved in regulation of cerebellar granule cell precursor cell proliferation	0.04976	0.03727	4/4	4/4	
Skeletal development						
GO:0035116	embryonic hindlimb morphogenesis	0.11589[§]	0.02289	15/28	18/28	47
GO:0048706	embryonic skeletal system development	0.01067	0.07440[§]	19/39	19/38	32
GO:0060348	bone development	0.01067	0.00589	26/38	29/38	
Metabolism						
GO:0050909	sensory perception of taste	0.00605	0.04735	14/31	14/31	
GO:0035634	response to stilbenoid	0.11589[§]	0.02248	5/6	6/6	50
GO:0051453	regulation of intracellular pH	0.04083	0.04516	6/7	6/7	
GO:0072001	renal system development	0.04632	0.03727	11/18	12/18	
GO:2000377	regulation of reactive oxygen species metabolic process	0.02304	0.03804	12/17	13/17	
Organismal development						
GO:0007275	multicellular organismal development	0.00165	0.00270	88/188	92/188	
GO:0009952	anterior/posterior pattern specification	0.00165	0.01467	41/87	44/87	
GO:0009953	dorsal/ventral pattern formation	0.03875	0.02248	23/43	27/43	
GO:0061154	endothelial tube morphogenesis	0.04868	0.04516	4/4	4/4	
Regulation						
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.01637	0.01467	181/460	217/460	
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.00165	0.00270	280/702	339/699	
GO:0030178	negative regulation of Wnt signaling pathway	0.01067	0.01467	24/42	26/42	

[a]*P*-value after adjustment for multiple testing; [b]the number of unique genes found for the given GO term related to the total number of genes that could be found at most for this term, i.e. genes that have a corresponding entry in the annotation file and contain at least one SNP; [c]for all non-significant GO terms (*P* FDR > 0.05), the rank of the term in the respective dataset (total terms $GO_{fixedSNPs}$: 5876, $GO_{PCAdapt}$: 6409); [§]GO term is statistically non-significant (*P* FDR > 0.05), but ranked among the top 50 terms

Fixed non-synonymous SNPs between species

Among the 123,023 SNPs fixed between Bornean and Sumatran orangutans, 296 SNPs were non-synonymous, i.e. altered amino acids, and three were splice donor/acceptor variants (Online Supporting Table SO8). A proportion of these SNPs likely represent causal variants underlying phenotypic differences between the two orangutan species. Fixed non-synonymous SNPs altered 236 protein-coding genes, of which 28 were uncharacterized novel genes in *Pongo* (identified by Ensembl Gene Build). Two fixed non-synonymous SNPs resulted in gain of a premature stop codon (loss-of-function mutations) in the genes *ARGFX* and *ZNF224*. *ARGFX* is a putative transcription factor and thought to be involved in early embryonic development. *ZNF224* may be involved in transcriptional regulation as repressor. We identified further loss-of-function mutations in splicing regions, which affected the gene *SRBD1* (splice donor variant), whose function remains unknown, and two uncharacterized novel genes (ENSPPYG00000010361, ENSPPYG00000000950). We did not find significant (*P* FDR < 0.05) enrichment of genes with potential functional changes for any particular biological GO term. Detailed functional and disease association information of all genes containing fixed non-synonymous SNPs or splice acceptor/donor variants are provided in Online Supporting Table SO10.

Seven of the genes with potential functional changes were located within putative selective sweep regions in Bornean orangutans, and nine within candidate regions of Sumatran orangutans (Tables 7 and 8). We have not modeled the functional consequences of the identified non-synonymous mutations on protein structure for these genes. Also, the exact biological function of most of these genes is currently unknown. However, at least three genes in Bornean orangutans (*KIAA1109*, *SPDL1*, *SMAD4*) were again associated with lipid and glucose metabolism (e.g. Croteau-Chonka *et al.* 2011). The gene *KIAA1109* was located within one of the 20 highest-ranking putative selective sweep windows (see above). For Sumatran orangutans, we recaptured the gene *TMEM43*, which was also located within the 20 top-ranking sweep windows and is related to heart function. Also associated with heart function is the gene *CD46* (Cho *et al.* 2009; den Hoed *et al.* 2013), though is also involved in immune response to pathogens. Three other genes in Sumatran orangutans (*SOX6*, *CTTNBP2*, and *GPSM2*) are among others involved in neurogenesis (Batista-Brito *et al.* 2009; Chen *et al.* 2012; Chen & Hsueh 2012; Doherty *et al.* 2012; Lee *et al.* 2014). The gene *GPSM2* also plays an important role in the development of hearing (Walsh *et al.* 2010; Doherty *et al.* 2012).

Table 5. List of fixed non-synonymous SNPs located within species-specific putative selective sweep regions.

Gene	Chromo	Position SNP	AA[a]	Codons	F_{ST}	d_{xy}	π_{PP}	π_{PA}	*P*-value[b]
P. pygmaeus									
TRMT10C	chr3	31,568,016	I/V	Att/Gtt	0.49874	0.00290	0.00033	0.00149	**
KIAA1109	chr4	126,951,069	S/T	aGc/aCc	0.52654	0.00211	0.00013	0.00108	***
IPO11	chr5	64,001,046	T/A	Aca/Gca	0.64024	0.00494	0.00052	0.00157	*
SPDL1	chr5	172,277,724	Q/P	cAa/cCa	0.55216	0.00258	0.00034	0.00108	*
FANCC	chr9	91,206,204	N/H	Aat/Cat	0.51587	0.00278	0.00030	0.00136	**
SMAD4	chr18	63,515,622	N/S	aAc/aGc	0.62305	0.00418	0.00035	0.00150	*
SMC1B	chr22	40,782,034	R/Q	cGa/cAa	0.52602	0.00291	0.00035	0.00135	*
		40,804,594	E/D	gaG/gaC					
P. abelii									
AHCTF1	chr1	2,121,249	S/N	aGt/aAt	0.50054	0.00262	0.00081	0.00092	*
CD46	chr1	42,344,323	G/R	Gga/Aga	0.49662	0.00289	0.00101	0.00093	*
GPSM2	chr1	119,345,368	R/K	aGa/aAa	0.51932	0.00279	0.00134	0.00050	**
EXOSC10	chr1	219,261,855	H/R	cAt/cGt	0.51061	0.00279	0.00084	0.00095	*
SRBD1[§]	chr2a	65,901,739	/	/	0.61086	0.00346	0.00063	0.00100	*
TMEM43	chr3	1,617,935	I/L	Att/Ctt	0.50944	0.00300	0.00173	0.00035	***
PEX1	chr7	83,653,328	V/I	Gtc/Atc	0.49174	0.00215	0.00079	0.00068	**
CTTNBP2	chr7	114,285,330	I/V	Atc/Gtc	0.59575	0.00340	0.00063	0.00105	*
SOX6	chr11	54,089,448	L/M	Ctg/Atg	0.52271	0.00219	0.00069	0.00068	**

[a]Altered amino acid; [b]Significance level of within-species π reduction (*: $P < 0.5$; **: $P < 0.1$; ***: $P < 0.01$); [§]Loss-of -function mutation

Table 6. Function of the genes in Table 5. A more detailed description of the genes is provided in the Online Supporting Table SO9.

Gene	Potential function
P. pygmaeus	
TRMT10C	functions in mitochondrial tRNA maturation; associated with hepatitis and malaria
KIAA1109	associated with susceptibility to celiac disease; potential role in adipocyte differentiation
IPO11	nuclear transport receptor for protein import
SPDL1	associated with adiposity; required for the localization of dynein and dynactin to the mitotic kinetochore, also required for correct spindle orientation
FANCC	DNA repair protein that may operate in a postreplication repair or a cell cycle checkpoint function; has been associated with body height in humans (Lango Allen *et al.* 2010)
SMAD4	regulates the transcription of target genes. E.g. positively regulates PDPK1 which has a central role in energy metabolism
SMC1B	Meiosis-specific component of cohesin complex
P. abelii	
AHCTF1	required for the assembly of a functional nuclear pore complex (NPC) on the surface of chromosomes as nuclei form at the end of mitosis
CD46	probably associated with heart rate; may also be involved in the fusion of the spermatozoa with the oocyte during fertilization; also involved in immune response to pathogens
GPSM2	may play a role in neuroblast division and in the development of hearing
EXOSC10	may participates in a multitude of cellular RNA processing and degradation events
SRBD1	unknown
TMEM43	transmembrane protein; defects cause arrhythmogenic right ventricular cardiomyopathy type 5
PEX1	required for the import of proteins into peroxisomes
CTTNBP2	controls dendritic spinogenesis in hippocampal neurons
SOX6	transcriptional activator; plays a key role in several developmental processes, including neurogenesis and skeleton formation

Between-species PCAdapt analysis

We further identified candidate SNPs for local adaptation based on PCA as implemented in the PCAdapt software. PCAdapt was able to capture previously characterized orangutan population structure (Figure 4, cf. Chapter 4). As expected, the first PC separated Bornean and Sumatran orangutans and the second PC Sumatran orangutans north of Lake Toba and Batang Toru to the south of it. In this exploratory study, we focused on selection along the axis of PC1, i.e. between Bornean and Sumatran orangutans. Gene ontology enrichment analysis of the 1% top-hit SNPs (n = 161,043; Online Supporting Table SO11) revealed 29 significant biological process GO terms (*P* FDR < 0.05; Online Supporting Table SO12), of which 13 overlapped with terms significant in the fixed-SNP analysis (Table 6).

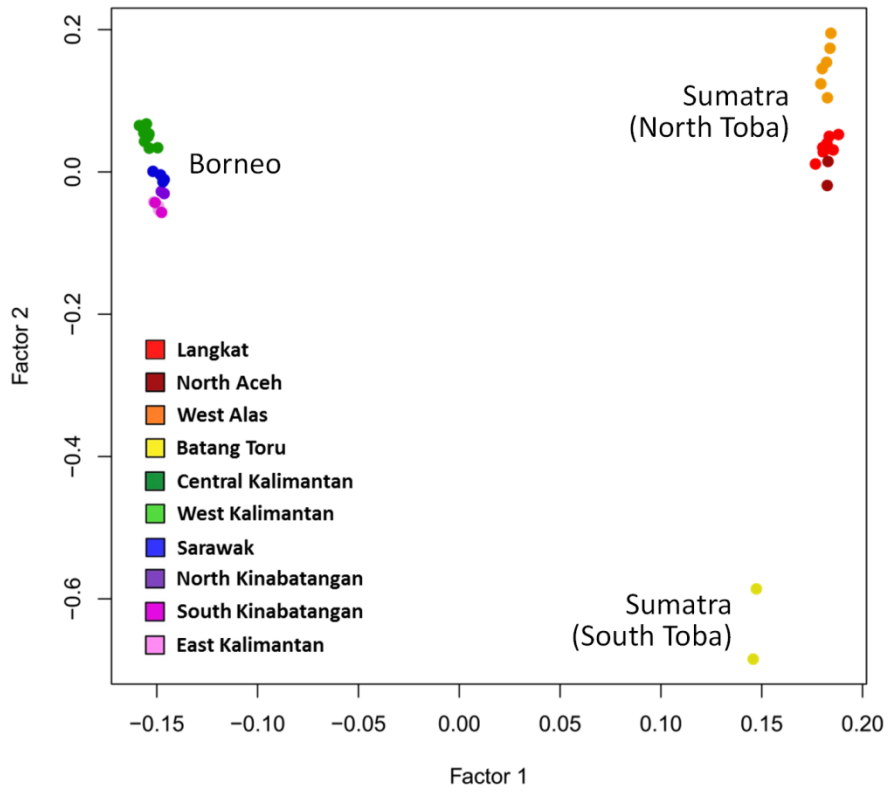


Figure 4. Principal component analysis of Sumatran and Bornean orangutans with PCAdapt. Each dot corresponds to one individual. Color codes match those of Figure 1.

PCAdapt analysis within Bornean orangutans

In light of the phenotypic differences between *P. p. wurmbii* and *P. p. morio*, we performed a second PCAdapt analysis exclusively within Bornean orangutans. PCAdapt correctly inferred extant orangutan population structure (Figure 5, cf. Chapter 4). The first PC separated orangutans of Central/West Kalimantan from the populations in Northeastern and Northwestern Borneo. The second PC distinguished the Sabah populations from the East Kalimantan and Sarawak populations. We focused on SNPs extensively correlated with the division axis of PC1, which largely corresponds to the gradient of environmental differences within Borneo.

Gene ontology analysis of genes co-located with the 0.5% highest-ranking SNPs (n = 53,122; Online Supporting Table SO13) revealed six significantly enriched GO terms (*P* FDR < 0.05; Table 7; Online Supporting Table SO14). We found an enrichment of genes involved in forebrain neuron development (Table 7). Furthermore, four GO terms were directly or indirectly associated with lipid and glucose metabolism. Many genes with high-ranking SNPs of the GO term 'regulation of cell size' are for instance involved in metabolism: *WDTC1* is associated with obesity, *RPTOR* encodes a component of a signaling pathway that regulates cell growth in response to nutrient and insulin levels, and *ATP2B2* plays a critical role in intracellular calcium homeostasis. In addition, most of the genes associated with the terms 'monoterpenoid metabolic process' and 'drug catabolic process' are members of the cytochrome P450 superfamily and catalyze reactions involved in synthesis of cholesterol, steroids, and other lipids. Expression of these genes is induced by glucocorticoids and starvation. Finally, all three genes associated with the GO term 'protein localization to lysosome' (i.e. *LAMTOR4*, *LAMTOR5*, and *SH3BP4*) act as indirect regulators of the mTORC1 nutrient-sensitive signaling complex that activates translation of proteins.

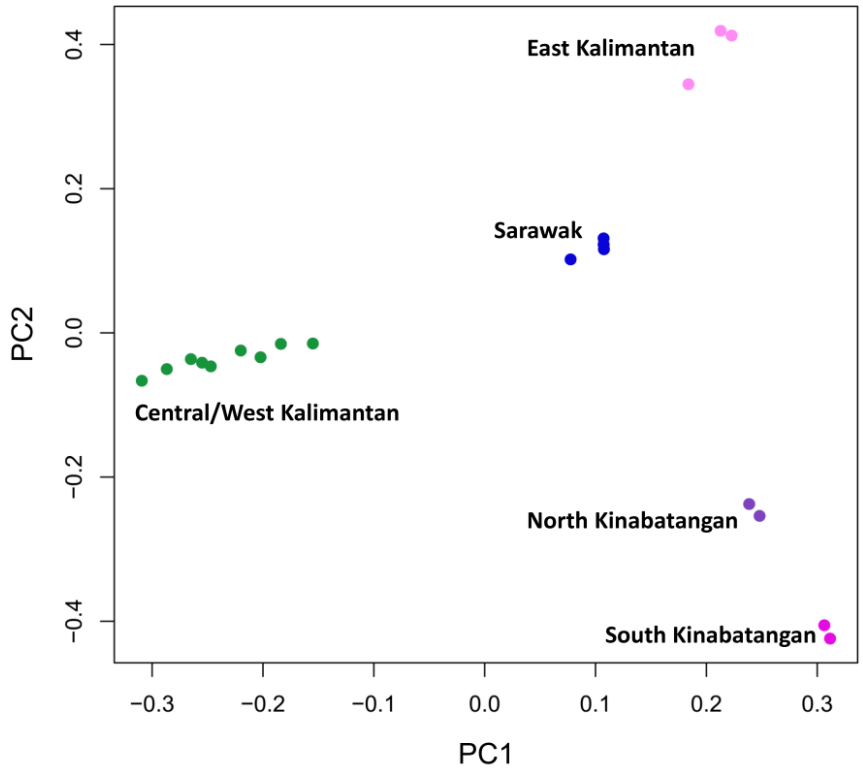


Figure 5. Principal component analysis of Bornean orangutans with PCAdapt. Each dot corresponds to one individual. Color codes match those of Figure 1.

Table 7. Gene ontology terms significantly enriched within Bornean orangutans. The tested gene set was derived from the 0.5% top-hit SNPs (Online Supporting Table SO13). We report only GO terms that are related to biological processes.

GO term	GO description	*P* FDR[a]	No. of genes[b]
GO:0008361	regulation of cell size	0.00455	6/9
GO:0040011	locomotion[§]	0.02706	6/8
GO:0061462	protein localization to lysosome	0.03068	3/3
GO:0021884	forebrain neuron development	0.03337	4/6
GO:0042737	drug catabolic process	0.03768	4/5
GO:0016098	monoterpenoid metabolic process	0.03768	3/4

[a]*P*-value after adjustment for multiple testing; [b]the number of unique genes found for the given GO term related to the total number of genes that could be found at most for this term, i.e. genes that have a corresponding entry in the annotation file and contain at least one SNP; [§]related to both movement of cells or whole organisms

6.5 Discussion

Our study is the first to extensively investigate the genetic basis of variation in phenotypic traits in a great ape genus, using whole-genome sequencing data of individuals representing all extant populations. Performing SNP and window-based genome scans, we identified a catalogue of candidate genes and functional variants potentially associated with genetic local adaptations in response to a west–east gradient in habitat productivity and stability of food supply.

Genetic adaptations in Bornean orangutans

Our results suggest for instance that genes related to lipid and glucose metabolism have evolved adaptively in Bornean orangutans. Several top-candidate genes within putative selective sweeps are involved in these processes (e.g. *PIK3R3*, *RABL3*, *SCAP*, *SLC6A2*, *KIAA1109*, *SPDL1*, and *SMAD4*). For some of the candidate genes (*KIAA1109*, *SPDL1*, and *SMAD4*), we identified potential causal target SNPs, i.e. fixed non-synonymous SNPs (between species). Moreover, the PCAdapt analysis between *P. p. wurmbii* and *P. p. morio* provided indication for selection on lipid and glucose metabolism also within Bornean orangutans.

Our findings are consistent with genetic adaptation in Bornean orangutans, especially *P. p. morio*, to cope with strong fluctuations of fruit abundance and unpredictable, prolonged periods of low energy intake in conjunction with ENSO events (Knott 1998; Delgado & van Schaik 2000; Wich *et al.* 2006; Morrogh-Bernard *et al.* 2009; van Schaik *et al.* 2009b). Most of the aforementioned candidate genes are directly or indirectly involved in energy storage (i.e. adipose tissue), which we link to potential changes in physiological buffering against starvation (Knott 1998; Morrogh-Bernard *et al.* 2009; van Schaik *et al.* 2009b; Isler 2014). This idea is supported by studies of physiology in wild orangutans and observations from captive animals, indicating that Bornean orangutans are better able at storing fat in adipose depots than Sumatran orangutans (Dierenfeld 1997; Knott 1998; Wich *et al.* 2006).

Apart from physiological changes, we also have indication for potential genetic adaptation associated with brain development in Bornean orangutans. Among the candidate genes within putative sweeps was for instance *FOXG1*, which has a crucial function in the development of the telencephalon and causes severe microcephaly (Hébert & McConnell 2000; Martynoga *et al.* 2005; Kortüm *et al.* 2011). In addition, we found that genes involved in forebrain pattern specification and in regulation of cerebellar granule cell precursor cell proliferation were significantly enriched in the analysis of fixed SNPs and in the between-species PCAdapt analysis. Also within Bornean orangutans, i.e. between *P. p. wurmbii* and *P. p. morio*, genes active in brain development (i.e. forebrain neuron development) were significantly overrepresented in GO analysis of PCAdapt top-hit SNPs. It must be stressed that the analyses of fixed SNPs and PCAdapt top-hit SNPs describe what differentiates orangutan

taxa at the single SNP level by either the impact of directional selection or also by pure genetic drift.

The described candidate genes and biological processes are associated with neurogenesis, which ultimately defines brain size by the number of produced neurons (Herculano-Houzel 2012; Lent *et al.* 2012). Though speculative, selective changes in these genes may build part of the genetic basis of the documented variation in brain size along the ecological west–east gradient, with the smallest brains found in the northeastern *P. p. morio* (Taylor & van Schaik 2007; C. P. van Schaik 2010, unpublished data). In line with the increased ability of fat storage, a decrease in brain size may represent adaptation to energy intake constraints during extended episodes of severe food scarcity (Taylor & van Schaik 2007; van Woerden *et al.* 2012). Selection might have favored reduction of brain size in order to reduce the costs of this metabolically expensive tissue (Rolfe & Brown 1997; Laughlin *et al.* 1998) in order to survive lean periods ("Expensive Brain framework", Isler & van Schaik 2009; van Woerden *et al.* 2012). Alternatively, the decrease in brain size could reflect a life history trade-off associated with a faster-paced life history in northeastern Bornean orangutans (van Schaik *et al.* 2009b).

Genetic adaptations in Sumatran orangutans

Due to differences in habitat, Sumatran orangutans do not have to cope with prolonged periods of food scarcity and have more favorable energy budgets than those on Borneo. In agreement with this and the documented differences in phenotypic traits, we identified a very different set of candidate genes for adaptive evolution in Sumatran orangutans.

One of our most intriguing findings was that we identified several genes within putative selective sweeps with crucial functions in learning, memory, and adult brain plasticity (e.g. *ATAD1*, *TBR1*, *TNPO1*, and *CDK6*). Selective changes in these genes may form part of a genetic basis providing Sumatran orangutans with a framework for extended flexibility (behavioral plasticity) for both individual learning, and social learning of local innovations ('culture') (van Schaik *et al.* 2003). Potentially also liked to their larger brains (van Schaik 2013), Sumatran orangutans show a larger cultural repertoire than Bornean orangutans (van Schaik 2004; van Schaik *et al.* 2009a; Krützen *et al.* 2011). Furthermore, while multiple complex innovations have been documented in Sumatran orangutans, similar complex innovations are rare to absent on Borneo (van Schaik *et al.* 2009a).

The higher levels of social learning in Sumatran orangutans are almost certainly closely associated with their higher sociability (Mitra Setia *et al.* 2009; van Schaik *et al.* 2009a; Weingrill *et al.* 2011). One of the top-ranking candidate genes of the window-based genome scan, *TNPO1*, plays a central role in the oxytocin pathway (Di Benedetto *et al.* 2014), which is known for stimulating social behavior (Kosfeld *et al.* 2005; Heinrichs *et al.* 2009; reviewed in Carter 2014). The functions of oxytocin are diverse and further include for instance the facilitation of lactation, maternal behavior, genetic regulation of the growth of the neocortex,

and maintenance of the blood supply to the cortex (reviewed in Carter 2014), all of which represent plausible targets for selection in Sumatran orangutans.

Beyond that, our results further suggest that genes relevant in heart development have evolved adaptively in Sumatran orangutans (top-candidate genes: *EPHA3*, *TMEM43*, and *CD46*). The potential underlying phenotypic targets are manifold, including genetic adaptation to increased energetic demands associated with their larger brains (Taylor & van Schaik 2007) and more active lifestyle, as for example deduced from the larger daily travel distances and higher frequency of day nests build (Singleton *et al.* 2009).

Furthermore, *GPSM2*, a gene playing a critical role in the development of hearing (Walsh *et al.* 2010; Doherty *et al.* 2012), may has been under positive selection in Sumatran orangutans. *GPSM2* is located within a strong putative selective sweep and contains a fixed functional variant (i.e. non-synonymous SNP). It is conceivable that this gene is associated with geographic variation in type and function of long calls emitted by flanged males and used for communication (Mitra Setia & Van Schaik 2007; Delgado *et al.* 2009; Spillmann *et al.* 2010; van Schaik *et al.* 2013).

Overall, for both orangutan species, our results suggest that an important mechanism for genetic adaptation may have been regulation of gene expression. For instance, several of the top candidate genes of the window-based genome scan were transcription regulators. Furthermore, only few protein-coding genes within putative selective sweep regions actually contained fixed potential functional SNPs (i.e. between-species fixed non-synonymous SNPs), implying that nearby regulatory elements may have been target of selection in many cases. Notably, we likely have also missed a number of functional SNPs due to the highly stringent filtering applied to identify fixed SNPs. We also found significant enrichment of genes for three GO terms directly related to regulation of transcription in the analysis of fixed SNPs and between-species PCAdapt top-hit SNP. An important role of gene expression regulation for local adaptation in orangutans would be consistent with data from humans and diverse other taxa suggesting that adaptation in regulatory elements is considerably more frequent than in protein-coding genes (e.g. Mikkelsen *et al.* 2007; Fraser 2013; Halligan *et al.* 2013; Konczal *et al.* 2015).

Methodological considerations and caveats

There are some important methodological considerations associated with this study. Alternative evolutionary processes can produce similar footprints in the genome as positive selection and it remains difficult to disentangle them (Jensen *et al.* 2005; Pavlidis *et al.* 2010). Although window-based genome scans can account to a certain degree for demographic factors (Excoffier *et al.* 2009; Bazin *et al.* 2010; Lawson *et al.* 2012), deriving significance thresholds at which the null hypothesis of neutral evolution can be rejected remains a challenging task (Crisci *et al.* 2012; Crisci *et al.* 2013). While we have good knowledge of the demographic history of the genus *Pongo* (e.g. Chapters 4 and 5; Nater *et al.* 2015), simulating

genomic regions under the inferred demographic model alone would be insufficient to derive significance thresholds, as long as the species-specific recombination rate variation is not included (O'Reilly *et al.* 2008; Auton *et al.* 2012; Roesti *et al.* 2012; Cruickshank & Hahn 2014). Currently, this information is not available for orangutans. In this study, we therefore applied an arbitrary 1%-cutoff for F_{ST}-outlier windows to identify potential candidates of selection without directly testing against the null model of neutral evolution. Furthermore, our analyses are based on the assumption that the genomic landscape of recombination is similar in both orangutan species, which is probably warranted considering their relatively recent divergence (Chapters 4 and 5; Nater *et al.* 2015).

Despite that we could not directly test against the null model of neutral evolution, by including measures of intraspecific genetic diversity, we could substantially reduce the false-positive error rate of putative selective sweeps in our F_{ST}-based windowed genome scans. We found that in 41% of the genomic regions with highly elevated F_{ST} (top 1% windows), genetic diversity was actually statistically significantly reduced in both orangutan species. We deem that genus-wide background selection, i.e. loss of genetic diversity in regions linked to sites under purifying selection (Charlesworth *et al.* 1993; Charlesworth 2013), likely caused the reduction of genetic diversity in most regions. Alternatively, genetic diversity may is reduced as result of independent parallel sweeps in both species, adaptive introgression (Hedrick 2013), or as a direct effect of reduced local recombination rate (Spencer *et al.* 2006). Because of the reduced intraspecific diversity, all of these processes inflate relative measures of differentiation such as F_{ST} (Charlesworth *et al.* 1997).

Aside from type-I errors, a major issue in studies of natural selection are also high false-negative rates (Crisci *et al.* 2013; Jensen 2014). The power to detect true selective sweeps depends on various factors, including the strength of selection, effective population size, time span, demographic history, population structure, and sample size (Olson-Manning *et al.* 2012; Crisci *et al.* 2013; Mita *et al.* 2013; Lotterhos & Whitlock 2014, 2015). We likely captured only a fraction of the genetic footprints of adaptation present in orangutan genomes. The absence of evidence is therefore not evidence of absence for positive selection acting on certain phenotypic traits varying within the genus *Pongo*. While this study mainly focused on the detection of recent hard sweeps, our whole-genome data will hopefully support further research on alternative mechanisms of genetic local adaptation, such as selection on standing genetic variation, which often results in soft selective sweeps.

In conclusion, the results of this study further our understanding of orangutan adaptive evolution and provide insights into what separates Bornean and Sumatran orangutans at the genomic level. Our findings suggest that at least some of the remarkable geographic variation in phenotypic traits in orangutans (compiled in Wich *et al.* 2009b) indeed represent unique genetic local adaptations. The identified candidate SNPs, genes and genomic regions provide a basis for more detailed examinations, integrating additional statistical tests and using larger sample sets of non-invasively collected samples from wild orangutan populations. Ultimately,

knowledge on the adaptive history of orangutans will shed light on how natural selection has shaped great ape genomes in general.

Acknowledgments

We thank David Marques for discussion on study design. For their help with collecting and exporting samples, we are grateful to Joko Pamungkas, Dyah Perwitasari-Farajallah, Muhammad Agil, and the staff at the Sumatran Orangutan Conservation Programme, Sepilok Orangutan Rehabilitation Centre, BOS Wanariset Orangutan Reintroduction Project, and Semongok Wildlife Rehabilitation Centre. Furthermore, we are indebted to numerous fellow orangutan researchers for gathering the data on the geographic variation in orangutan behavioral ecology that made the comparisons of this study possible. In addition, we thank the following institutions for supporting our research: Sabah Wildlife Department (SWD), Indonesian State Ministry for Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), Leuser International Foundation (LIF), Taman National Gunung Leuser (TNGL), and Borneo Orangutan Survival Foundation (BOSF). Major financial support was provided by the UZH University Research Priority Program, Leakey Foundation (to MPG), ERC Starting Grant (grant no. 260372 to TMB), Swiss National Science Foundation (grant no. 3100A-116848 to MK and CPvS), Forschungskredit University of Zurich (to MPG), Julius–Klaus Foundation (to MK), A.H. Schultz Foundation (to MK and MPG), and the Anthropological Institute & Museum at the University of Zurich.

Author Contributions

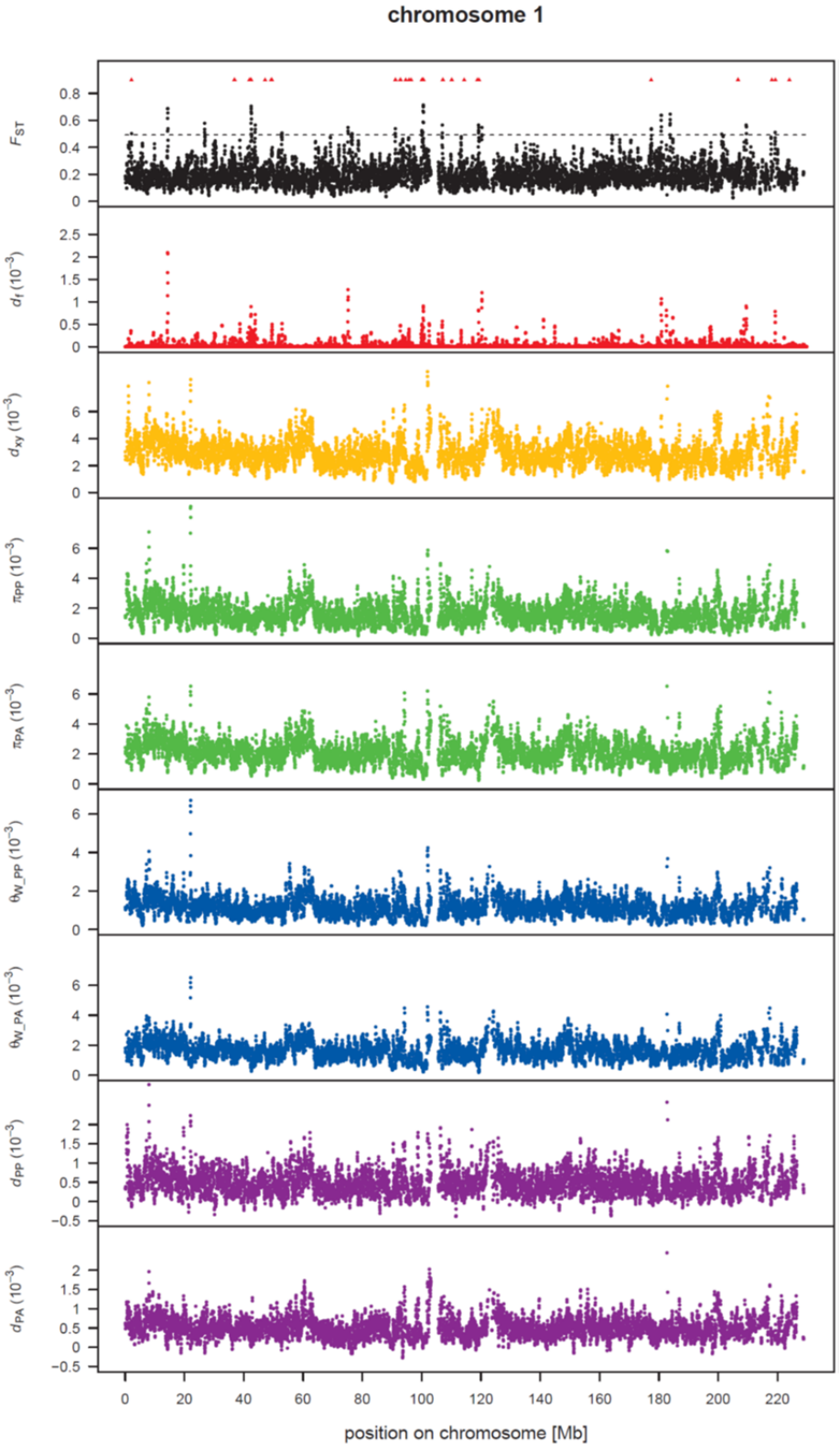
MPG, AN, and MK conceived the study. BG, MPG, MK, EV, KS, IS, AN, LNA, and CPvS provided genetic samples. IG and MG carried out sequencing. TMB and JPM contributed additional sequencing data. AN and MPG performed short read mapping, SNP and genotype calling. MPG designed and conducted statistical analyses. AN contributed to design of statistical analyses and provided novel bioinformatic tools. MPG wrote the manuscript. AN and MK edited the manuscript.

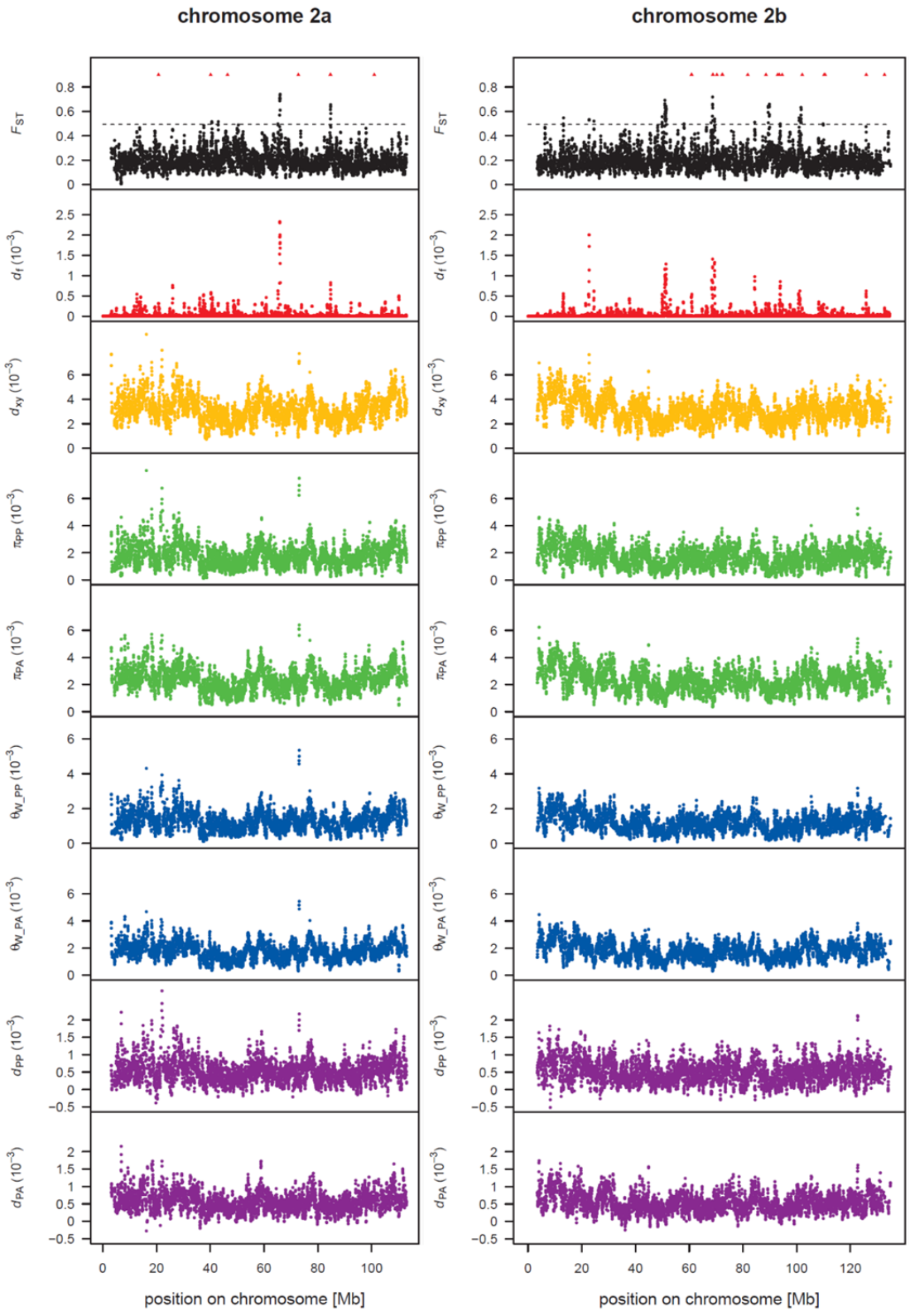
6.6 Supporting Information

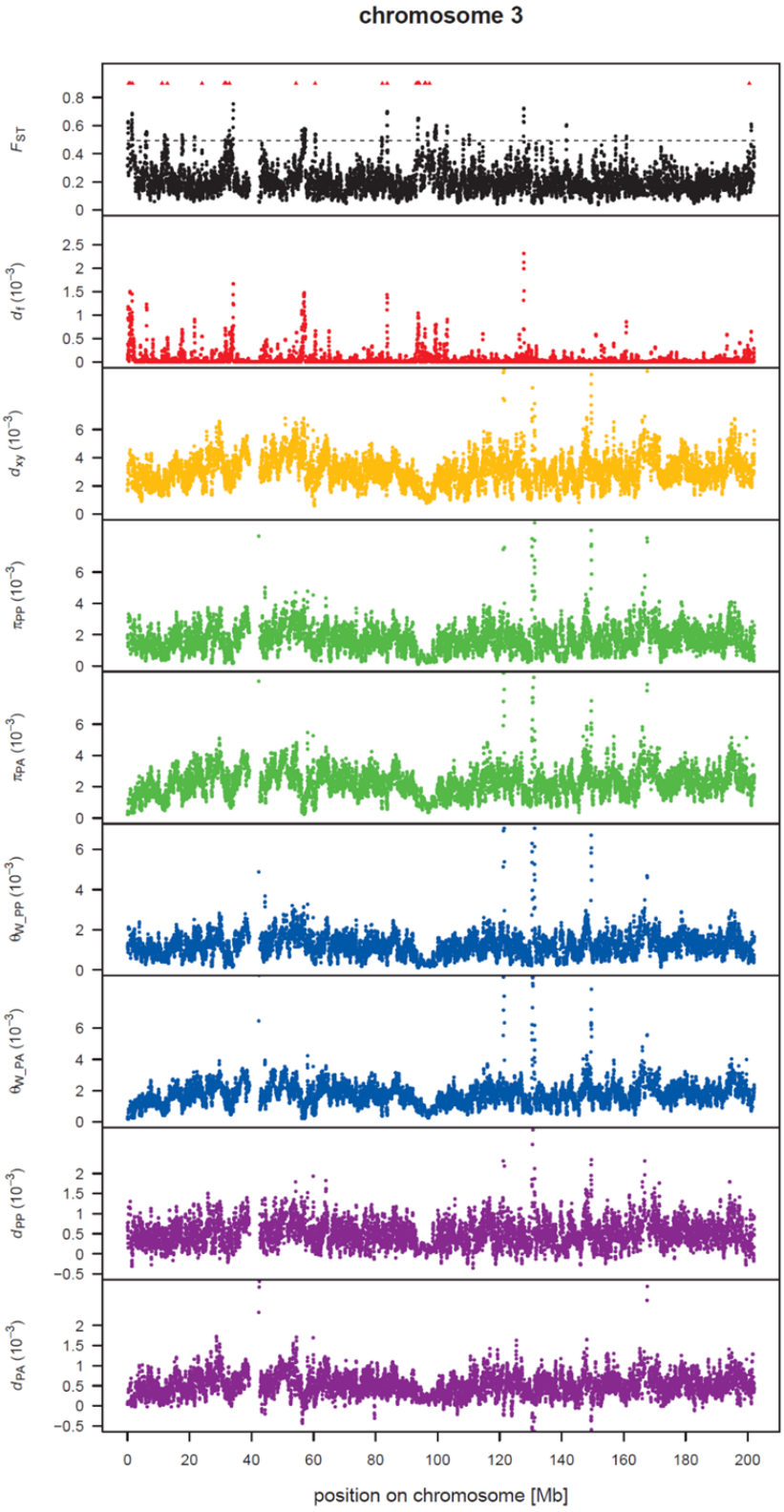
The Online Supporting Information will be deposited on the Dryad Digital Repository and is currently available from here:

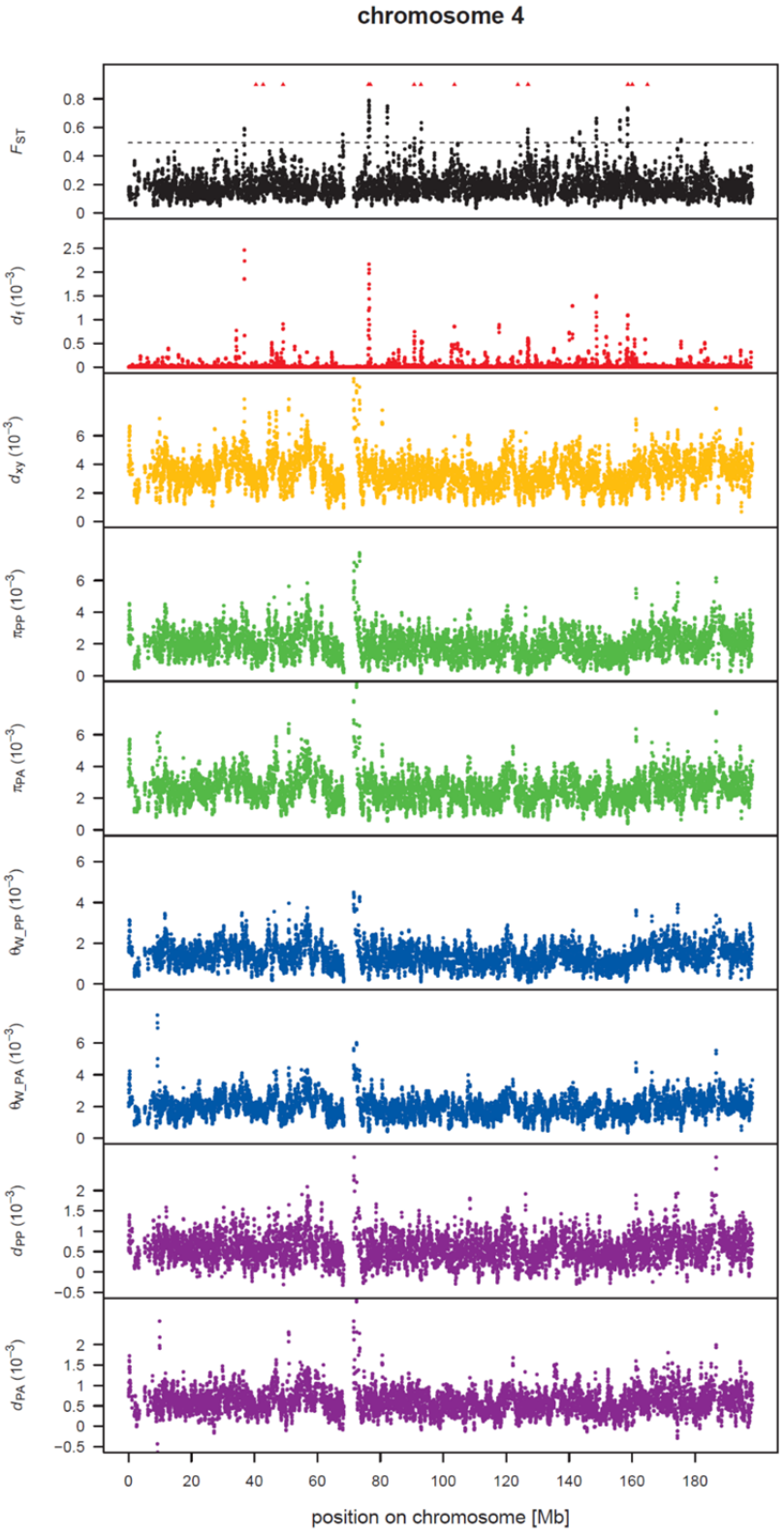
https://www.dropbox.com/s/6n6vnanj6fdiq1r/Greminger_2015_Thesis_Chapter-6_Online-Supporting-Information.xlsx?dl=0

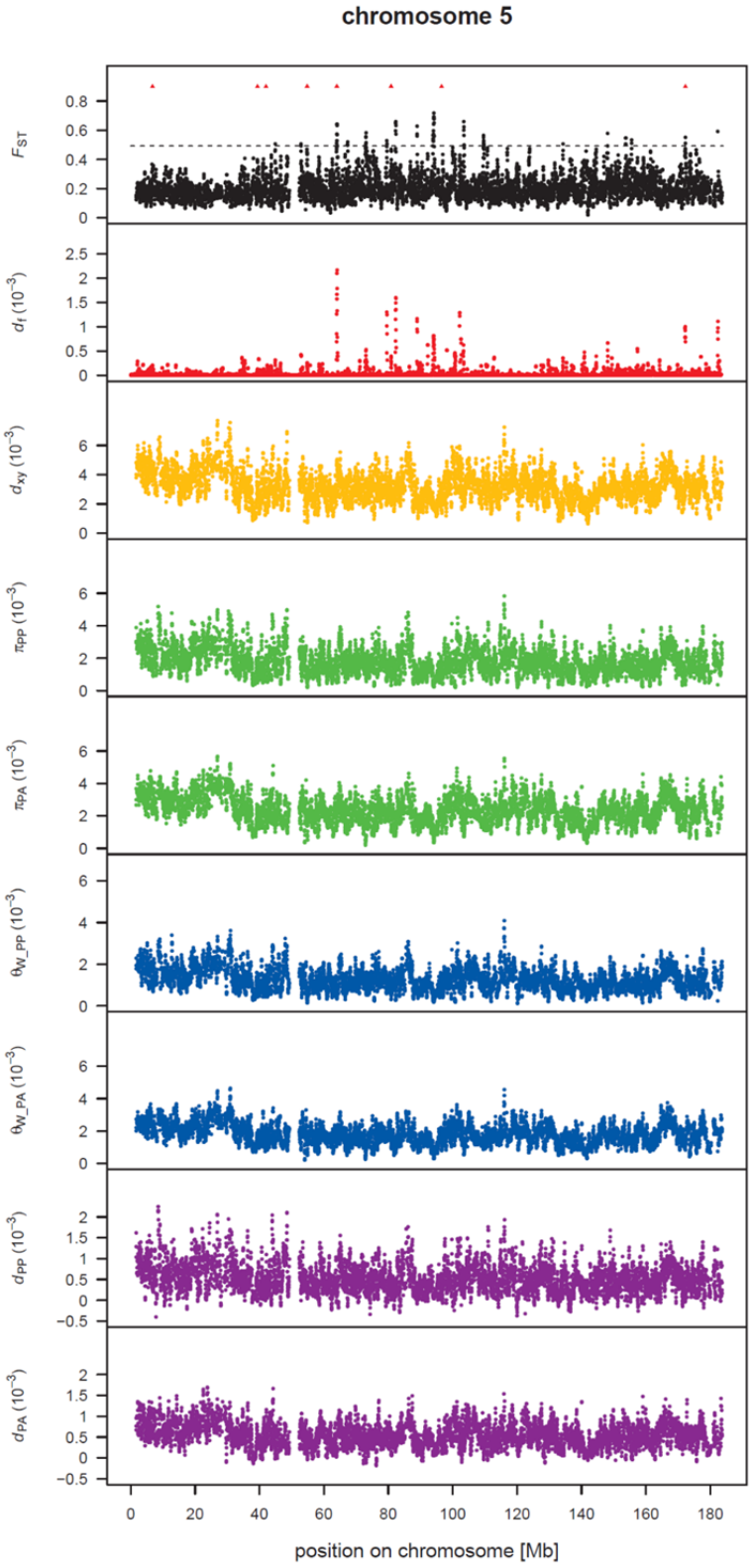
Supporting Figure S1. Distribution of windowed population genetic summary statistics along chromosomes 1–22. Summary statistics (y-axis) were averaged in windows of 100 kb length, sliding in 25 kb steps along the chromosome (x-axis). Plotted are the between-species population differentiation (F_{ST}), the density of fixed differences between species per base pair (d_f), the mean pairwise between-species sequence divergence (d_{xy}), the within-species nucleotide diversity for Bornean (π_{PP}) and Sumatran orangutans (π_{PA}), the within-species Watterson estimator ($\theta_{W\_PP/PA}$), and within-species Tajima's d ($d_{PA/PP}$). The dashed black line indicates windows above the 99th percentile of the empirical F_{ST} distribution. The small red triangles at the top of the F_{ST} plot denote chromosome positions of fixed non-synonymous SNPs between species.

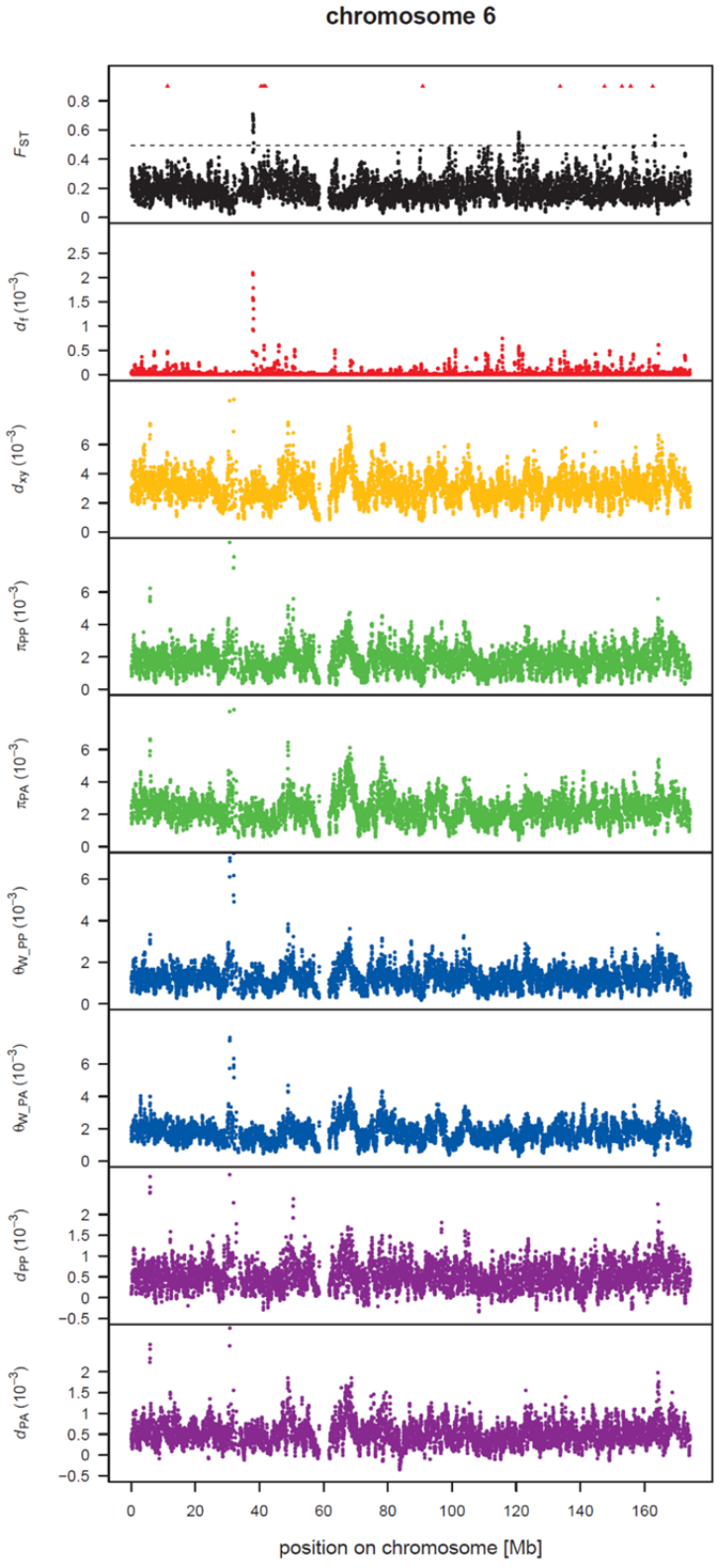

chromosome 1
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140 160 180 200 220
position on chromosome [Mb]

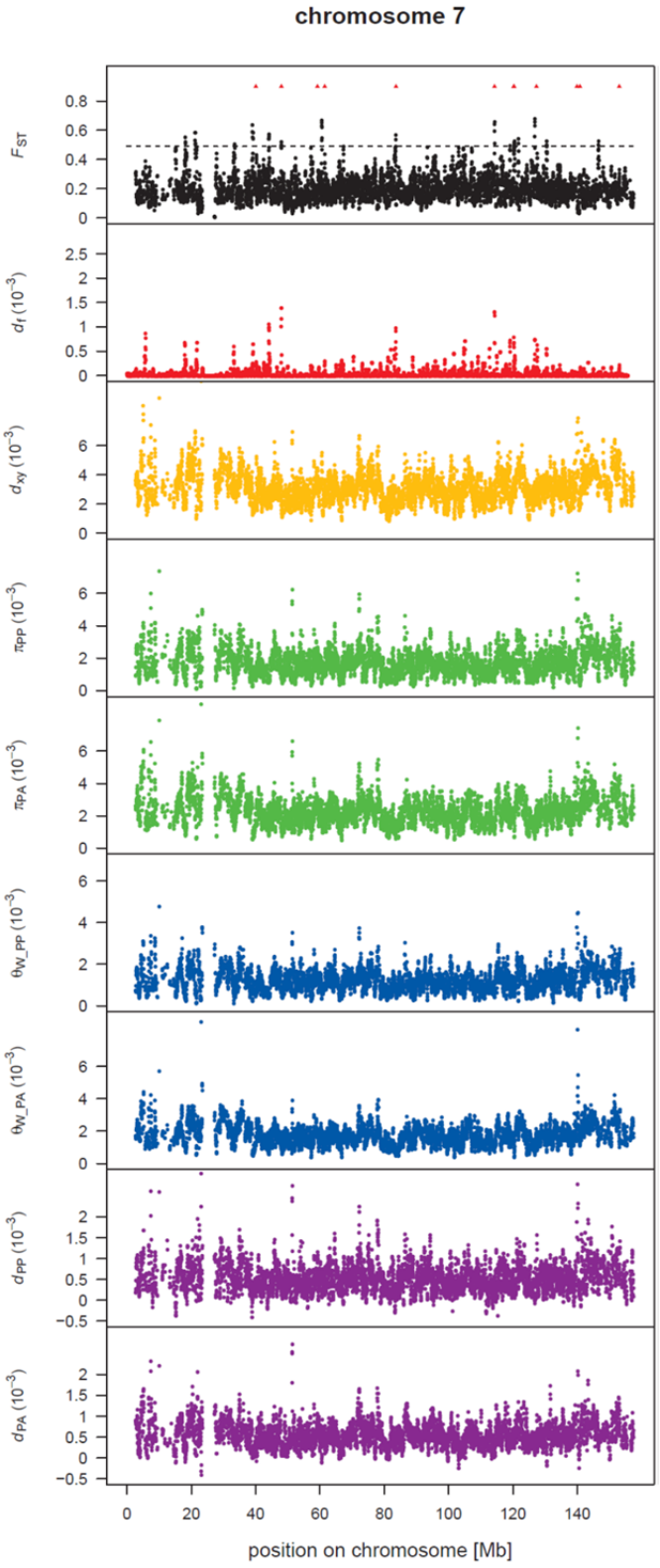

chromosome 2a
chromosome 2b
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]

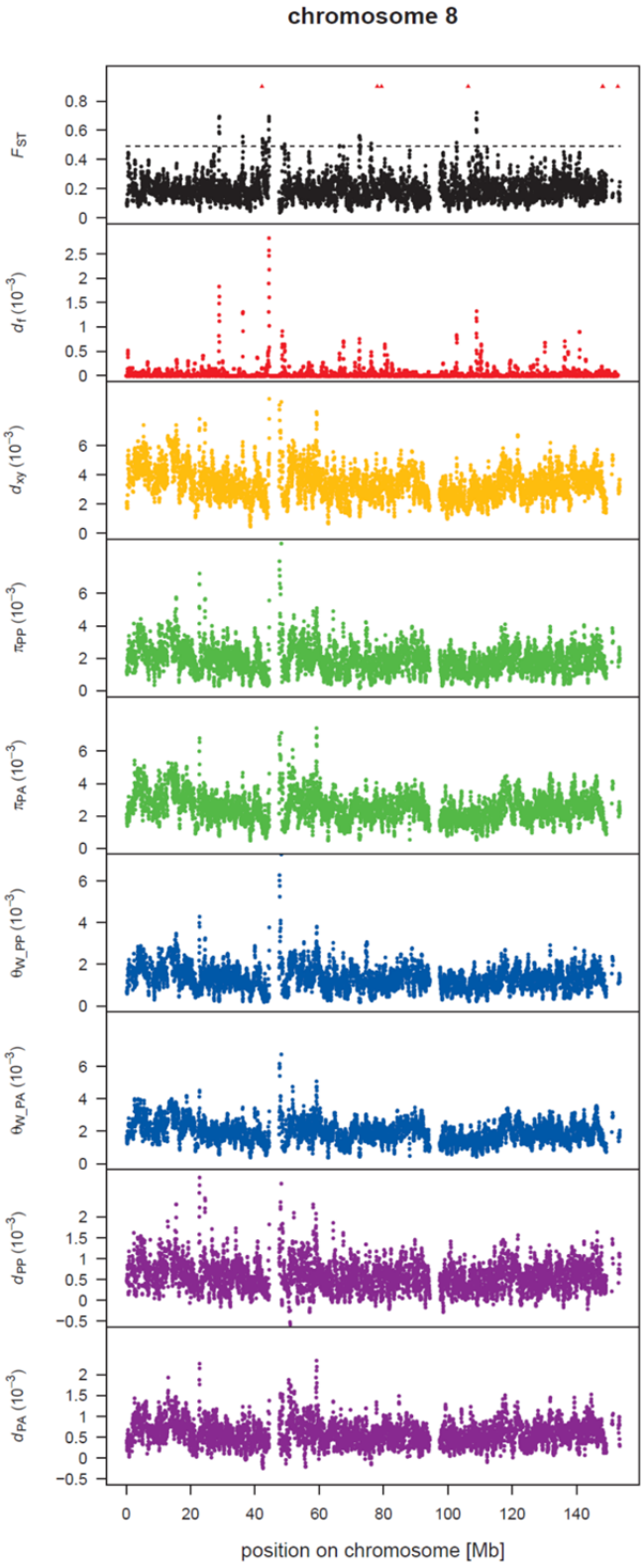

chromosome 3
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140 160 180 200
position on chromosome [Mb]

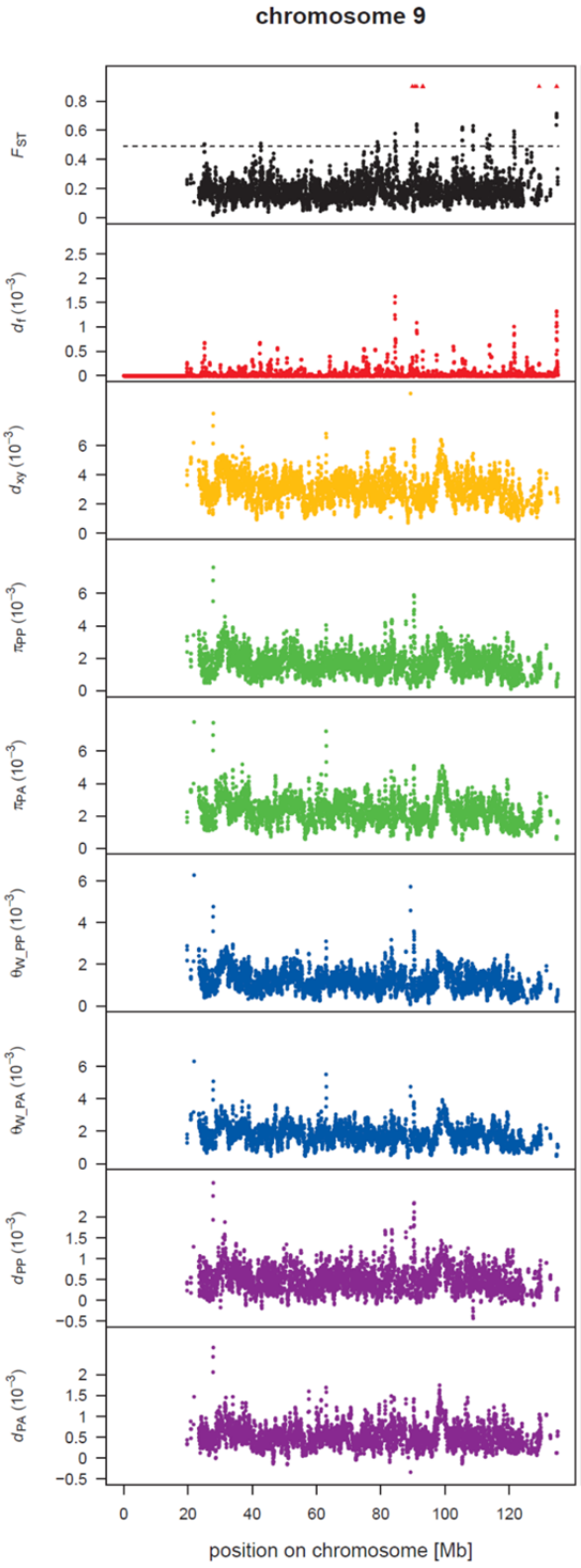

chromosome 4
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140 160 180
position on chromosome [Mb]

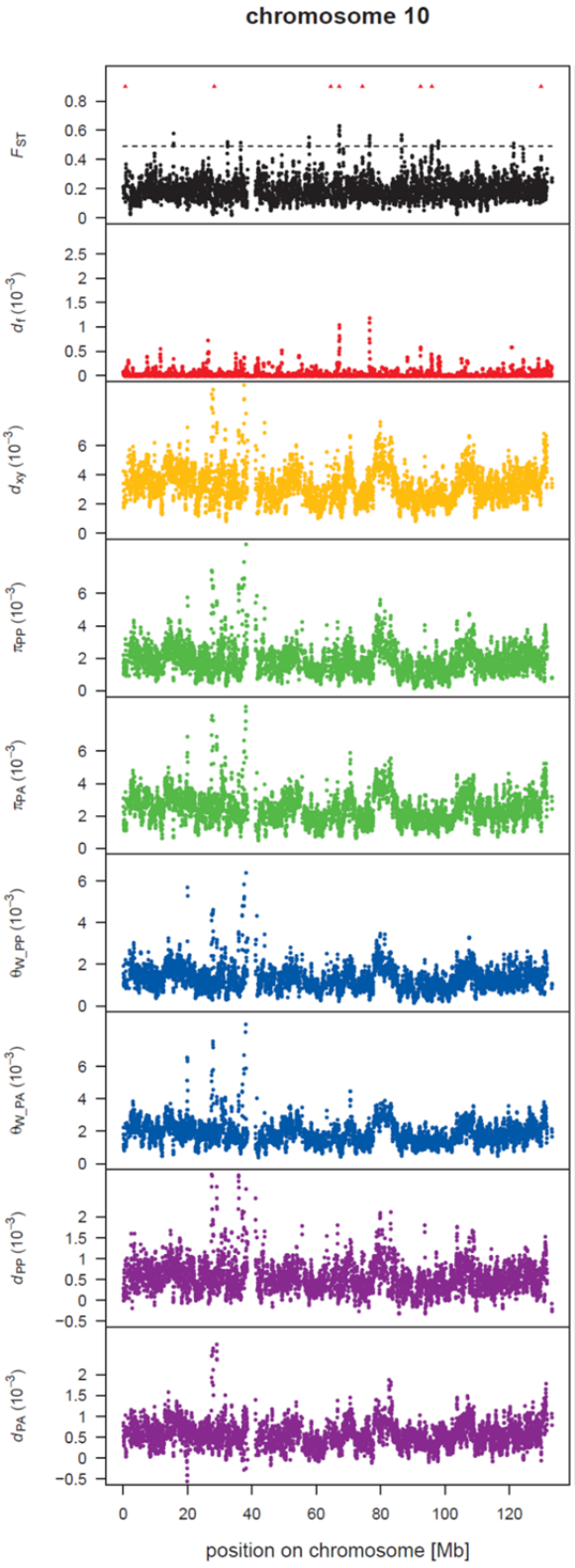

chromosome 5
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140 160 180
position on chromosome [Mb]

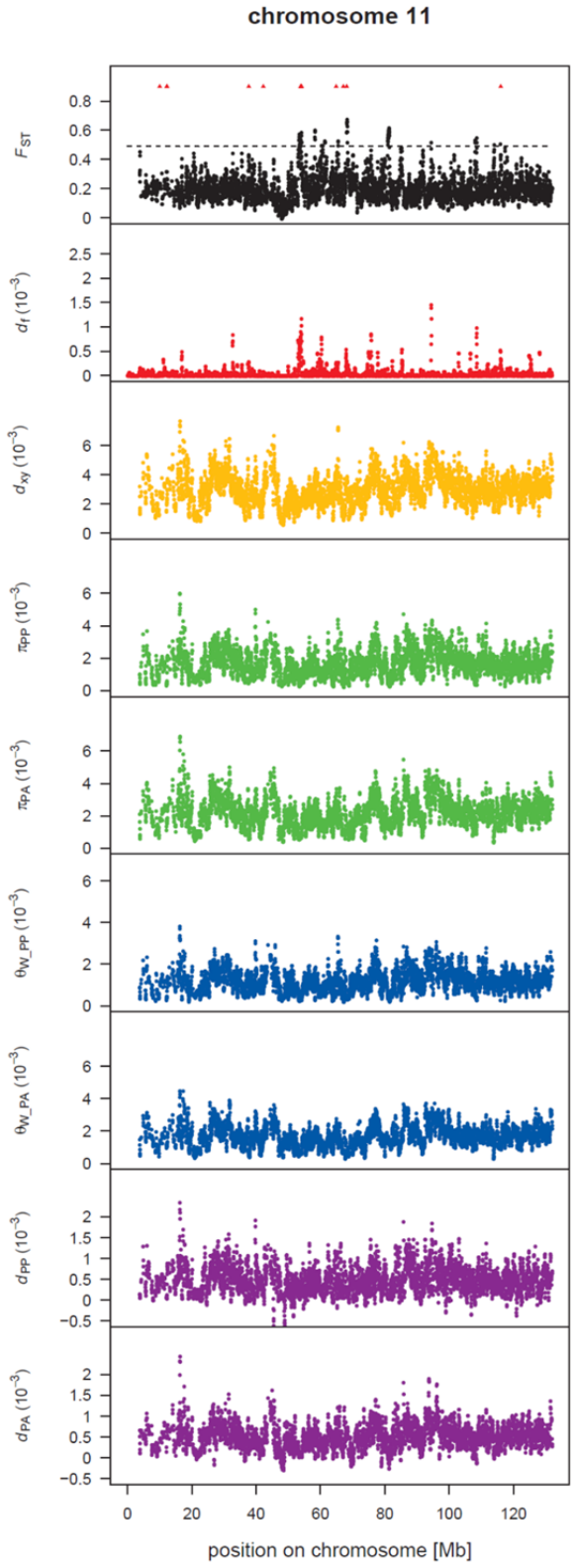

chromosome 6
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140 160
position on chromosome [Mb]

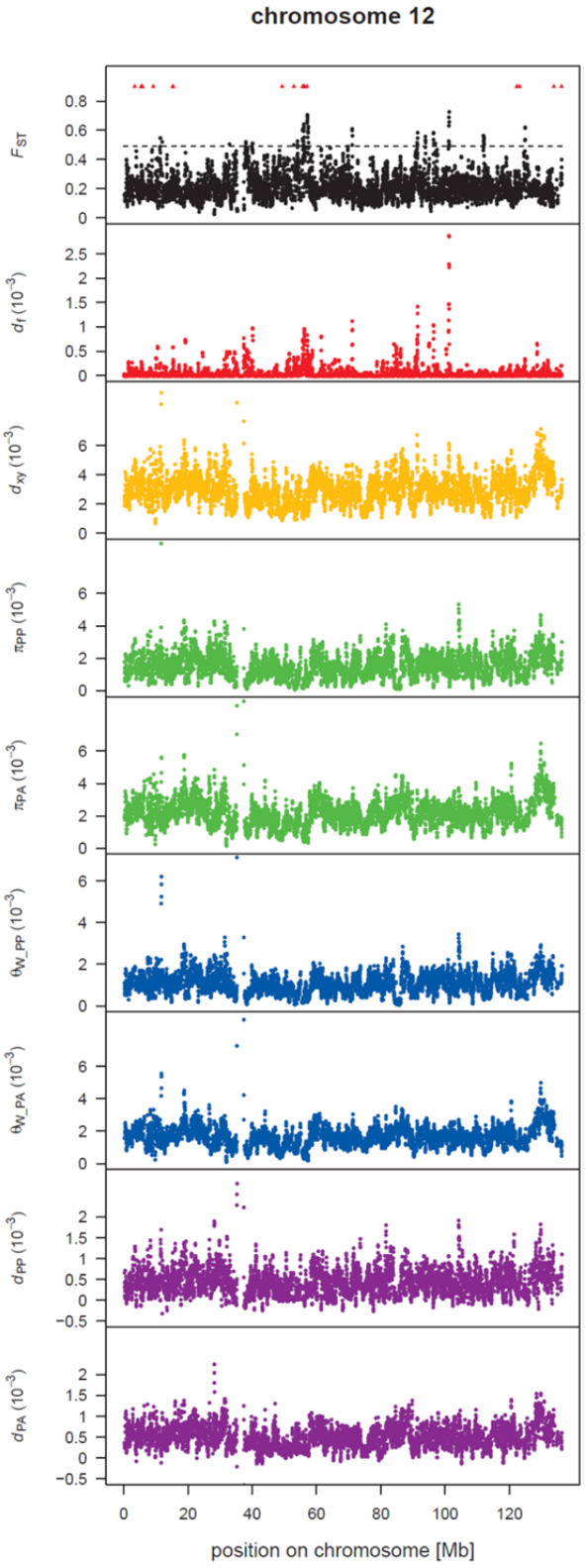

chromosome 7
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140
position on chromosome [Mb]

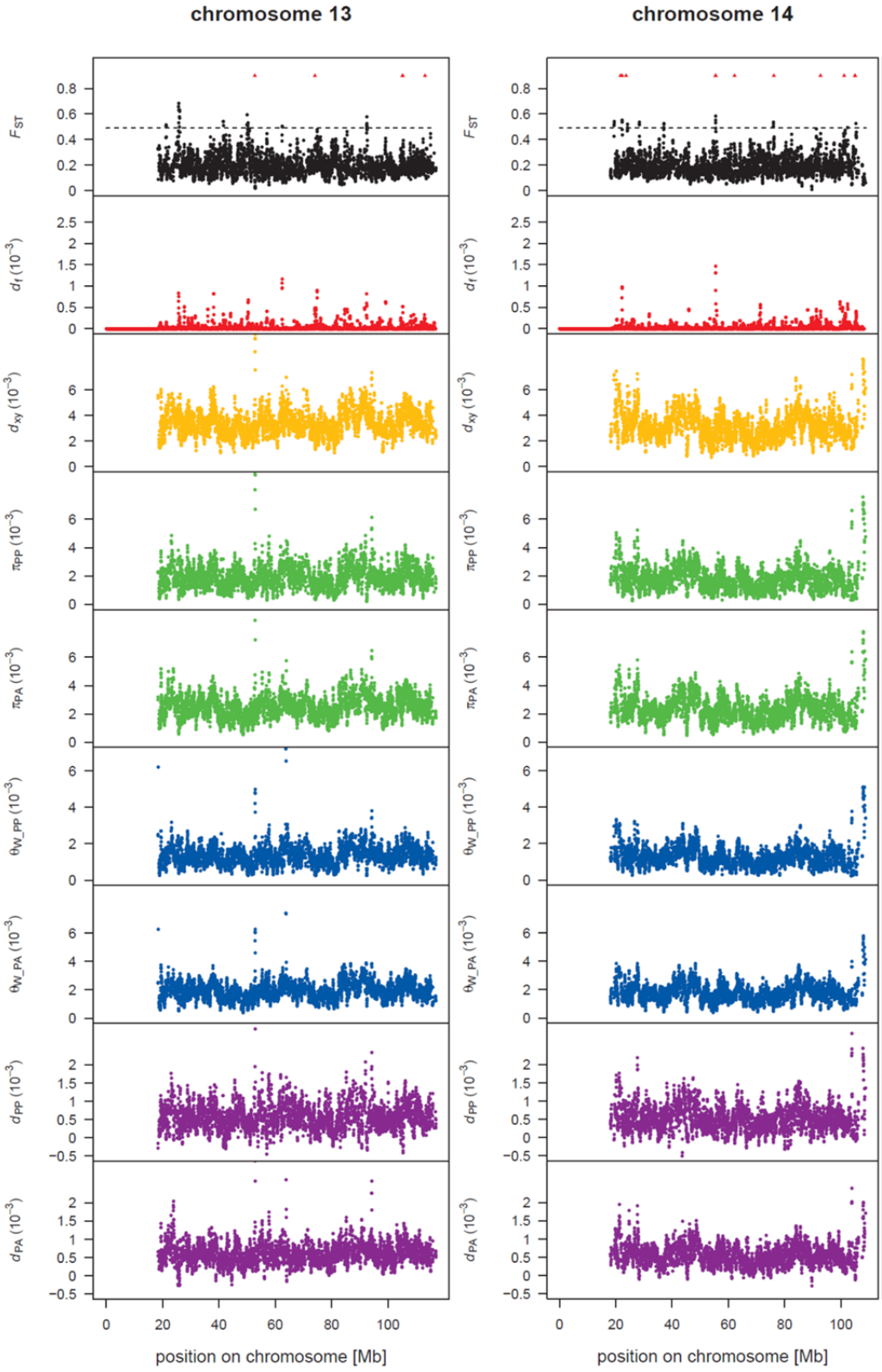

chromosome 8
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140
position on chromosome [Mb]

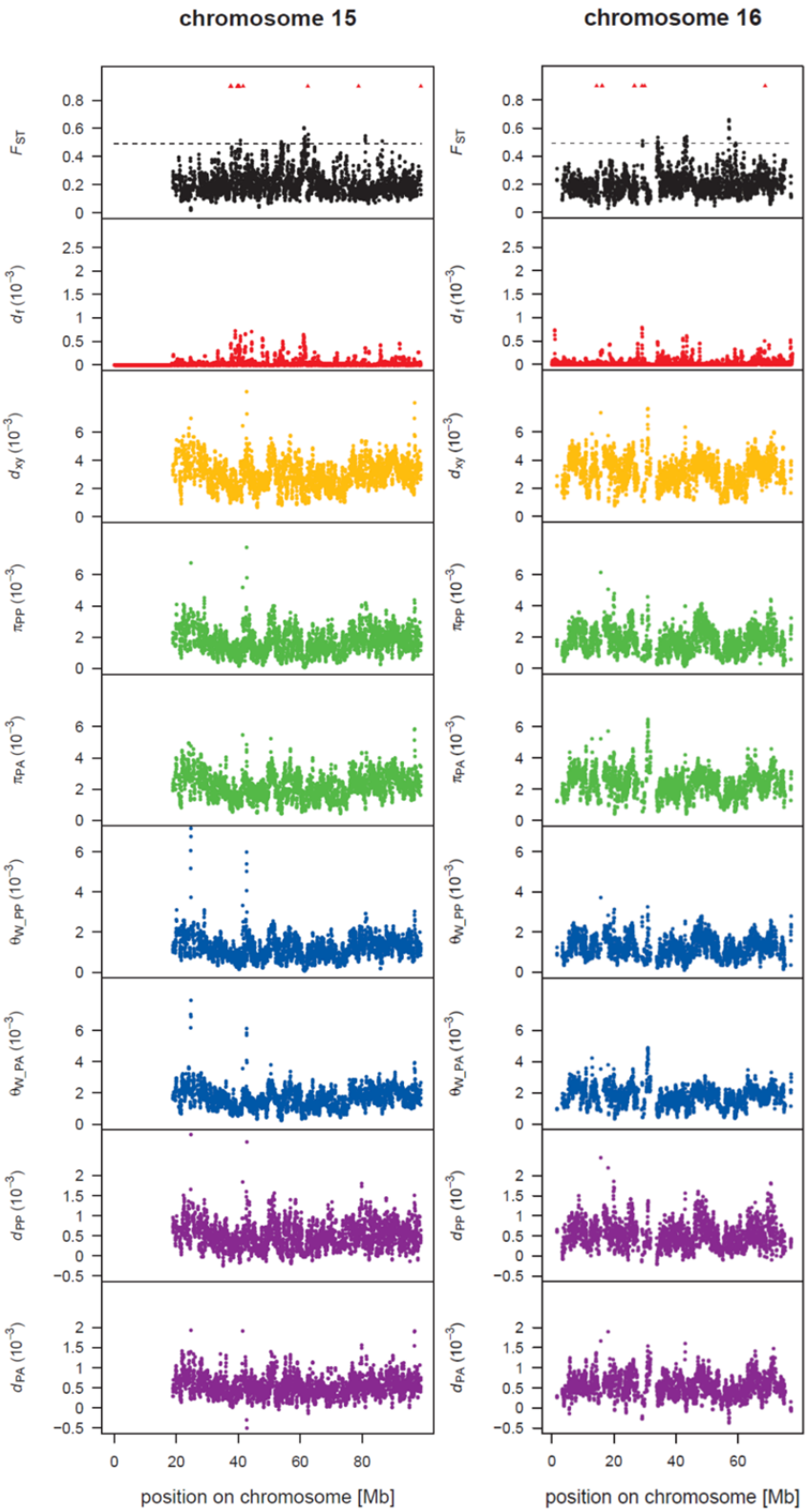

chromosome 9
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]

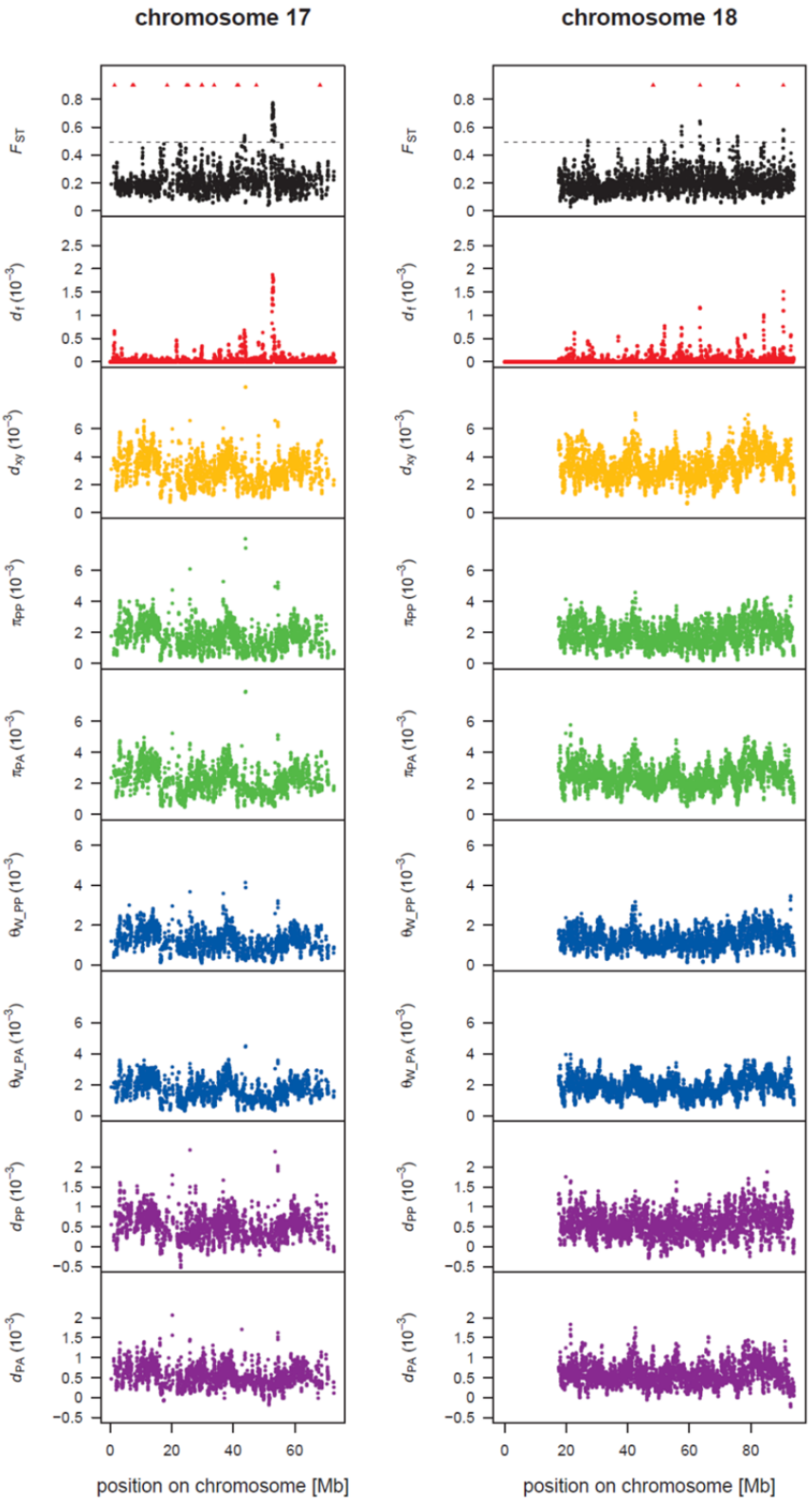

chromosome 10
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120
position on chromosome [Mb]

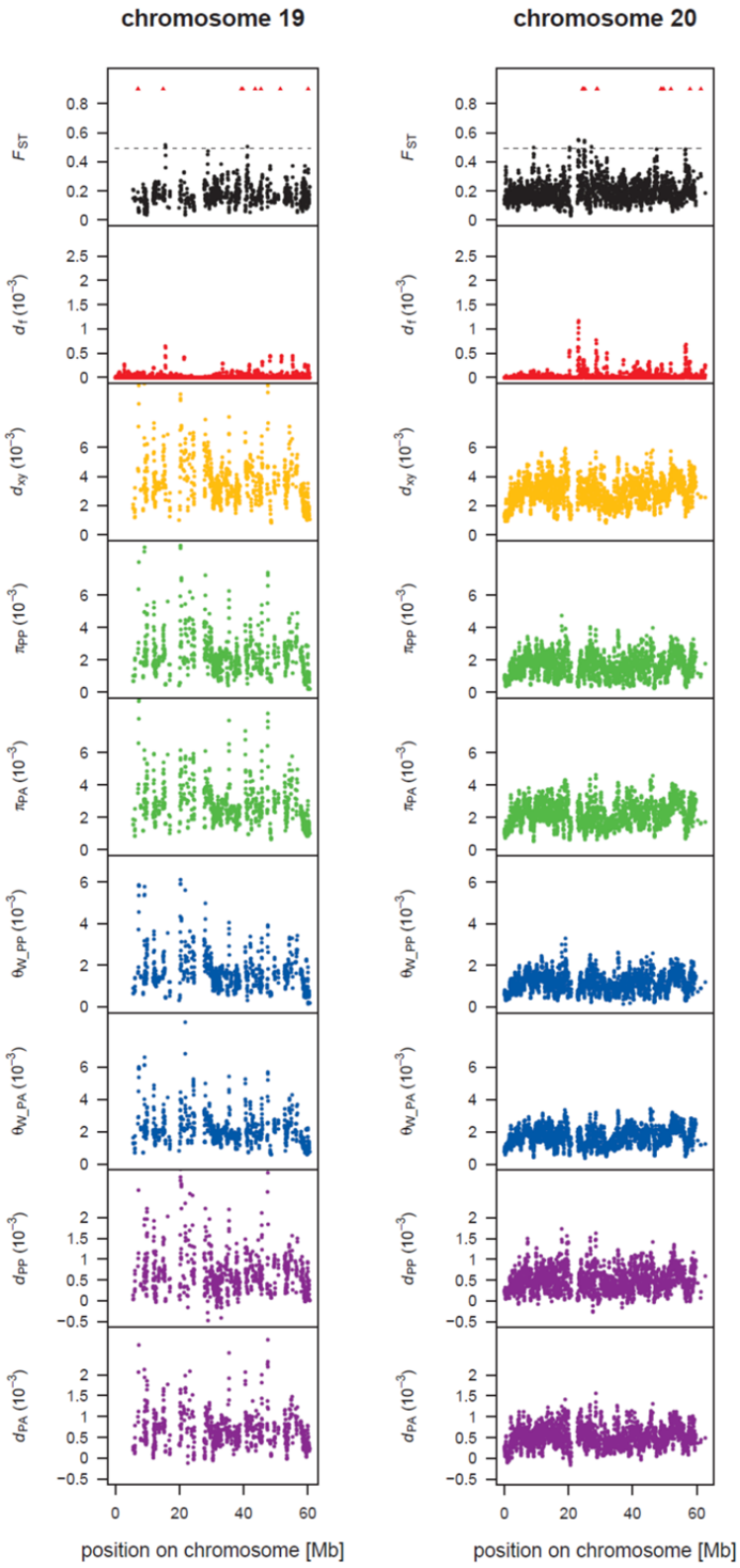

chromosome 11
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120
position on chromosome [Mb]

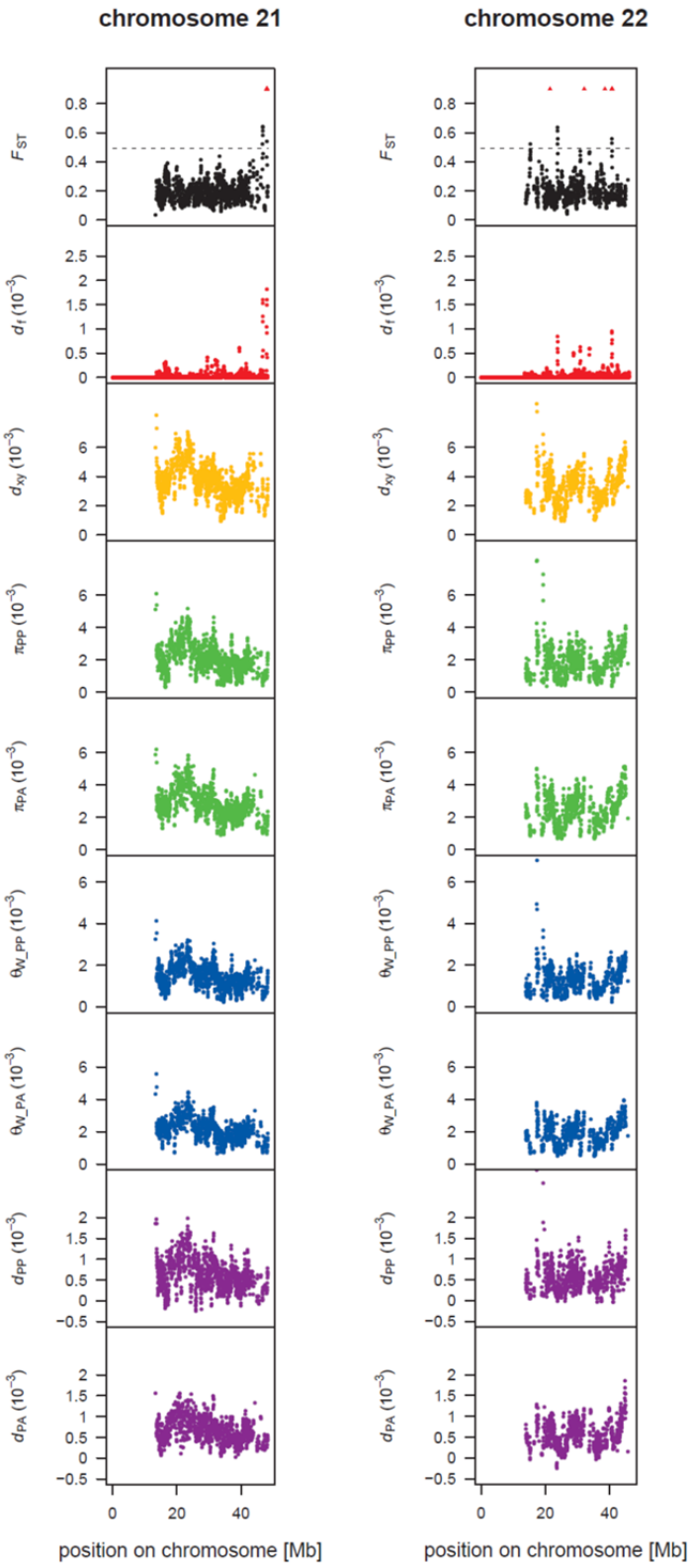

chromosome 12
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120
position on chromosome [Mb]

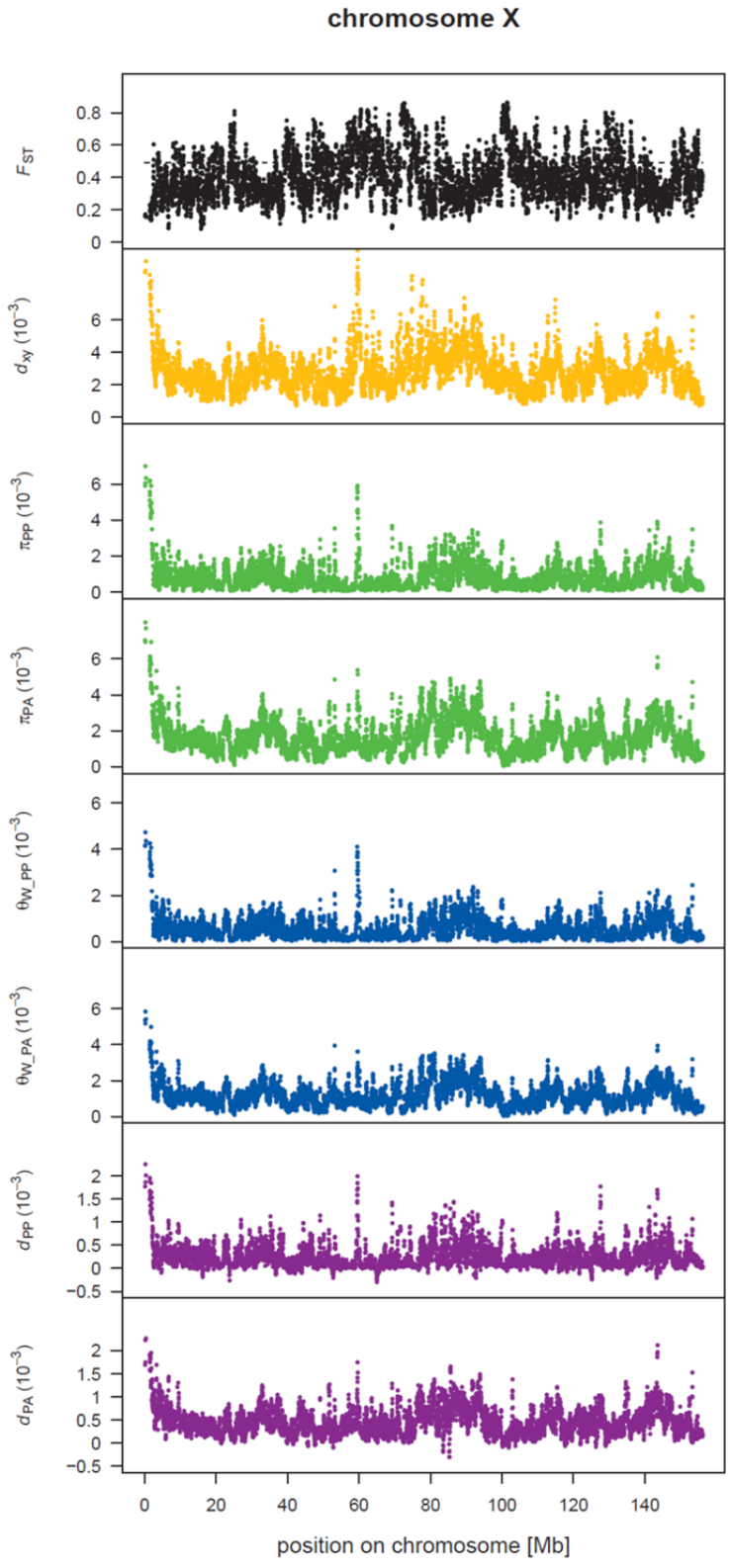

chromosome 13
chromosome 14
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]


chromosome 15
chromosome 16
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]


chromosome 17
chromosome 18
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]


chromosome 19
chromosome 20
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]


chromosome 21
chromosome 22
F_{ST}
d_f (10^{-3})
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
position on chromosome [Mb]
position on chromosome [Mb]


chromosome X
F_{ST}
d_{xy} (10^{-3})
π_{PP} (10^{-3})
π_{PA} (10^{-3})
θ_{W_PP} (10^{-3})
θ_{W_PA} (10^{-3})
d_{PP} (10^{-3})
d_{PA} (10^{-3})
0 20 40 60 80 100 120 140
position on chromosome [Mb]

Chapter 7

General discussion and perspectives

The advent of high-throughput sequencing has opened up new avenues in the study of evolutionary genetics. This dissertation takes advantage of this revolution when studying population genomics of orangutans with respect to the highly dynamic environmental processes taking place in Sundaland. In this chapter, I provide a synthesis of the main conclusions about the evolutionary history of the genus *Pongo* (Chapters 4–6) and the implications for taxonomy and conservation. Finally, I will outline some perspectives for future work. The main findings of the methodologically focused Chapters 2 and 3 have been summarized in the methodological outline of this dissertation (Chapter 1.4) and will not be reviewed here again.

7.1 The evolutionary history of the genus *Pongo*

Orangutans experienced a complex evolutionary history, which was impacted by the highly dynamic environmental conditions on the Sunda archipelago, and the pronounced sex-biased dispersal in this genus. Although previous genetic and genomic studies offered important insights into the complex demographic history and phylogeography of orangutans, they also came up with conflicting results. Central aspects remained poorly understood. Previous work has suffered from either the use of a small number of genetic markers (e.g. Arora *et al.* 2010; Nater *et al.* 2011), or from unknown population provenance of samples, small sample size, and missing MSY-specific data (e.g. Locke *et al.* 2011; Prado-Martinez *et al.* 2013). The adaptive evolutionary history of different orangutan taxa and the genetic basis underlying the remarkable variation in phenotypic traits (van Schaik *et al.* 2009b) have largely remained unexplored. In this dissertation (Chapters 4–6), I studied the evolutionary history of the genus *Pongo*, i.e. how environmental processes in Sundaland shaped patterns of genetic variation, and the genetic basis underlying local adaptations, by applying the most comprehensive genomic sampling to date. Together with my colleagues, I generated and analyzed a unique dataset of autosomal whole-genome data, complete mitochondrial genomes, and large-scale MSY-specific data of orangutans representing the entire extant geographic distribution of the genus (detailed in Chapters 4 and 5). It is the first time that orangutans have been studied at the genome-wide level using samples with known population provenance.

Colonization of the Sundaland islands

Mitochondrial genome data (Chapter 5) indicate that the ancestors of extant orangutans colonized the islands of Sundaland in the late Pliocene by entering Sumatra from the Southeast Asian mainland. According to the split sequence of the clades in the *Pongo* mtDNA phylogeny (Chapter 5) and considering potential migration routes from the mainland based on paleogeographical reconstructions of Sundaland (Hall 2013; de Bruyn *et al.* 2014), orangutans most certainly colonized central Sumatra initially and subsequently expanded to the north (i.e. north of present-day Lake Toba). The division axis around Lake Toba corresponds to the oldest spit in the orangutan mtDNA phylogeny with a TMRCA ~3.5–4.0 Ma (Chapter 5). Borneo seems to have been colonized in the early Pleistocene from

southern/central Sumatra, as indicated by the basal position of the Batang Toru mtDNA lineage to that of all extant Bornean orangutans with a TMRCA ~2.0–2.5 Ma (Chapter 5). My mtDNA results (tree shape and split times) are well in line with a previous study (Nater *et al.* 2011) that only used three mitochondrial genes, suggesting that the use of mitochondrial genome data improve mtDNA phylogenies in old species with extreme female philopatry only marginally.

Orangutan speciation

We found that the speciation of Bornean and Sumatran orangutans has been a gradual process over several hundred thousand years, and appears to have been heavily influenced by recurrent climate changes and high levels of male-biased dispersal and strict female philopatry (Chapters 4 and 5; Nater *et al.* 2015). Genomic data of both the autosomes (Chapter 4) and the MSY (Chapter 5) provide evidence that after the initial separation of Bornean and Sumatran orangutans at the beginning of the Pleistocene, autosomal gene pools remained connected via regular male-mediated gene flow over the cyclically exposed Sunda Shelf. Autosomal gene pools of both islands finally diverged ~0.9–1.1 Ma (Chapter 4), indicating a substantial reduction of levels of gene flow around that time. This reduction in migration opportunities may has been associated with a fundamental change in Earth's climate system at the transition from Early to Middle Pleistocene (Head & Gibbard 2005; Elderfield *et al.* 2012).

The more recent coalescent time of orangutan MSY lineages at ~430 ka (Chapter 5) indicates that periodical male migration between Borneo and Sumatra still has continued for another few hundred thousand years until gene flow finally completely ceased. The inferred cessation of gene flow between orangutan species is considerably earlier than proposed by studies using classical genetic markers, which suggested ongoing gene flow as recent as the penultimate (190–130 ka) or even last (110–18 ka) glacial period (Muir *et al.* 2000; Verschoor *et al.* 2004; Steiper & Young 2006; Becquet & Przeworski 2007; Nater *et al.* 2011; Nater *et al.* 2015). Our estimate, however, is compatible with findings by Locke *et al.* (2011) and Mailund *et al.* (2012) based on autosomal genome data, suggesting substantially reduced levels of gene flow between species at ~300-400 ka.

Our MSY data indicate that habitat conditions on the exposed landmass at low sea levels during the last few glacial periods of the Pleistocene must have prevented orangutans, and therefore likely also other forest-dwelling species, to disperse over the shelf. This may have been associated with an increasingly drier and seasonal climate, and speaks strongly against extensive lowland rainforest coverage on the exposed shelf as hypothesized by some paleoecological reconstructions (Sun *et al.* 2000; Cannon *et al.* 2009; Wang *et al.* 2009). A broad savanna corridor and large river systems dissecting the exposed Sunda Shelf (Rijksen & Meijaard 1999; Bird *et al.* 2005; Harrison *et al.* 2006; Slik *et al.* 2011) probably imposed impassable barriers to dispersal for rainforest-dwelling species. Taken together, our results point at complex temporal fluctuations of levels of gene flow between Bornean and

Sumatran orangutans, most likely related to varying climate among and within Pleistocene glacial periods (Cannon *et al.* 2009; de Bruyn *et al.* 2014).

The population history of Bornean orangutans

Following their divergence, Bornean and Sumatran orangutans experienced drastically different population histories (Chapters 4 and 5; Nater *et al.* 2015). As detailed in the Chapters 4 and 5, Bornean orangutans most probably underwent several population bottlenecks during the Pleistocene, likely caused by repeated extensive rainforest contractions and expansions associated with the climate oscillations (Flenley 1998; Morley 2000; Bird *et al.* 2005). The recent coalescent times of Bornean mitogenome and MSY lineages (Chapter 5) provide strong support for the late-Pleistocene refugium hypothesis (Arora *et al.* 2010; also see Nater *et al.* 2011; Nater *et al.* 2015), proposing that Bornean orangutans were confined within a rainforest refugium during the penultimate glacial period (130–190 ka), which has been particularly harsh (Martinson *et al.* 1987; Wright 2000). The basal position of the two Sabah populations (North and South Kinabatangan) to all other Borneans in our mitogenome phylogeny (Chapter 5) is congruent with the idea that this common glacial refugium was located in the Crocker mountain range in northern Borneo, as proposed previously for the Sabah orangutans (Jalil *et al.* 2008). The expansion from the late-Pleistocene refugium led to a only recently established population structure on Borneo (Chapters 3–5; Arora *et al.* 2010; Greminger *et al.* 2014; Nater *et al.* 2015).

Our results from the autosomal genome (Chapter 4) suggest that the expansion from this refugium was not paralleled with a large and stable increase in effective population size. On the contrary, our inferences revealed Bornean orangutans having experienced a population decline over the past ~300,000 years, resulting in very low effective population sizes in the more recent past. This long-term decline is likely associated with change to a drier and more seasonal climate, which probably also has been causal for the cessation of gene flow between the islands (discussed above). The Toba supereruption ~73 ka (Chesner *et al.* 1991) may have further decimated Bornean orangutans. Unfortunately, however, because their population size was already very small we did not have the resolution to detect a potential signal with the method we applied (Chapter 4).

The population history of Sumatran orangutans and the Toba supereruption

In contrast to Bornean orangutans, the autosomal effective population size of Sumatran orangutans actually appears to have increased during the Middle Pleistocene (~1–0.1 Ma; Chapter 4). Furthermore, as evident from the deep divergence of mtDNA lineages (Chapter 5; Nater *et al.* 2011), the population structure of Sumatran orangutans has been remarkably stable throughout the Pleistocene. Overall, I found that Sumatran orangutans were much less affected by the Pleistocene climate oscillations than the Borneans (Chapters 4 and 5). This is most likely due to the different geology and environmental conditions of Sumatra which

provided relatively stable rainforest coverage or at least multiple glacial refugia) along the Barisan Mountain range during glacial periods (Gathorne-Hardy *et al.* 2002).

Our genomic data (Chapters 4 and 5) provide strong support, however, for a major impact of Mount Toba on the evolutionary history of Sumatran orangutans. The extensive activity of this supervulcano (Chesner *et al.* 1991) caused a deep separation of orangutan populations to the south and to the north of it (Chapters 4 and 5; Nater *et al.* 2011; Nater *et al.* 2013). This finding bears important ramifications for the conservation and taxonomy of orangutans (outlined in section 7.5). Moreover, our inferences from autosomal genomes (Chapter 4) disclosed a drastic collapse of the effective population size of Sumatran orangutans coinciding with the Toba supereruption ~73 ka (Chesner *et al.* 1991) from which orangutans did not manage to recover. This lack of demographic recovery might be attributed to prehistoric hunting by early humans, which is believed to have been responsible for the complete disappearance of orangutans in many areas of Sundaland in the Late Pleistocene to early Holocene (Rijksen & Meijaard 1999; Delgado & van Schaik 2000).

Our results add to the controversial discussion about the consequences of the Toba supereruption. Although the Toba supereruption ~73 ka was the most powerful explosive eruption of the Quaternary (Chesner *et al.* 1991; Rampino & Ambrose 2000), its impact on global climate and ecosystems remains highly debated (e.g. Schulz *et al.* 2002; Gathorne-Hardy & Harcourt-Smith 2003; Petraglia *et al.* 2007; Haslam & Petraglia 2010; Williams *et al.* 2010; Williams 2012). It has been hypothesized that the eruption induced a 'volcanic winter' that, among others, may have caused a severe population bottleneck in early humans (Ambrose 1998; Rampino & Ambrose 2000). Gathorne-Hardy & Harcourt-Smith (2003) argued that under such a scenario, one would expect at least a similar population decimation of regional, environmentally sensitive taxa. To our knowledge, our results provide the first genetic evidence of such a strong regional impact on a large mammal. The deep divergence of Sumatran orangutan populations (Chapter 5; Nater *et al.* 2011), however, shows that local orangutan populations on Sumatra did not went completely extinct and thus that rainforest had not been destroyed over vast areas as advocated by some researchers (Rampino & Ambrose 2000; Williams *et al.* 2009).

Genetic local adaptations in orangutans

Our whole-genome scans for positive selection (Chapter 6) revealed that Bornean and Sumatran orangutans experienced very different adaptive evolutionary histories and that at least some of their striking geographic variation in phenotypic traits (Wich *et al.* 2009b) may indeed represent genetic local adaptations. Their adaptive evolutionary histories have likely been greatly influenced by their differential population histories and were probably mostly shaped by differences in overall forest productivity, the extent of El Niño-Southern Oscillation phenomenon (ENSO) impact, and consequences of climate changes.

Our results suggest that Bornean orangutans, in particular those in the northeast of the island (*P. p. morio*), may have genetically adapted to cope with strong temporal fluctuations of fruit abundance and prolonged periods of low energy intake associated with unpredictable ENSO events and Pleistocene climate oscillations (Knott 1998; Delgado & van Schaik 2000; Wich *et al.* 2006; Morrogh-Bernard *et al.* 2009; van Schaik *et al.* 2009b). Several of the genes within top-candidate selective sweep regions identified in Bornean orangutans are for instance involved in energy storage (i.e. adipose tissue) metabolism, which is consistent with the greater ability of Bornean orangutans to deposit large fat storages compared to Sumatran orangutans (Dierenfeld 1997; Knott 1998; Wich *et al.* 2006). These metabolic changes are assumed to allow for physiological buffering against starvation (Knott 1998; Morrogh-Bernard *et al.* 2009; van Schaik *et al.* 2009b; Isler 2014).

Furthermore, we also detected signatures of potential genetic adaptation associated with brain development, in particular neurogenesis, which defines brain size by the number of produced neurons (Herculano-Houzel 2012; Lent *et al.* 2012). This finding is compatible with the smaller brain size of the northeastern Bornean orangutans (Taylor & van Schaik 2007; C. P. van Schaik 2010, unpublished data) and may again represent an adaptation to survive prolonged lean periods by reducing the costs of this metabolically expensive tissue ("Expensive Brain framework", Isler & van Schaik 2009; van Woerden *et al.* 2012). Alternatively, it may reflect a life history trade-off associated with their faster-paced life history (van Schaik *et al.* 2009b).

It is tempting to speculate about interesting parallels between the adaptive evolutionary history of *P. p. morio* and the enigmatic Homo floresiensis from the Sundaland island of Flores, who is characterized by its usually small stature and remarkably reduced brain size (Brown *et al.* 2004; Falk *et al.* 2005). It is still matter of debate whether H. floresiensis indeed represents a new hominin species, or a pathological form of modern human (e.g. Henneberg *et al.* 2014). It has been hypothesized that the special characteristics of H. floresiensis ("insular dwarfism") represent adaptations to energy intake shortages (Brown *et al.* 2004), which would imply that the genus Homo is "morphologically more varied and flexible in its adaptive responses than previously thought" (Brown *et al.* 2004). Drawing parallels to *P. p. morio*, it seems plausible that adaptation to ecological factors account for the special characteristics of H. floresiensis, who probably had to cope with the same environmental constrains as *P. p. morio*, in particular with severe impacts of the ENSO (Taylor & van Schaik 2007). Undoubtedly, it would be highly interesting to explore the genetic basis of potential convergent evolution. Unfortunately, however, such an endeavor will be highly limited by several factors, including the quality and quantity of ancient DNA from tropical regions. In fact, attempts to extract DNA from teeth samples of Homo floresiensis have failed (Jones 2011).

In contrast to Bornean orangutans, orangutans from northern Sumatra have little exposure to extended periods of food scarcity (Wich *et al.* 2006; Morrogh-Bernard *et al.* 2009; Wich *et al.* 2011b). Also forest productivity is overall higher in Sumatra (Husson *et al.* 2009; Marshall *et*

al. 2009; Wich *et al.* 2011b). Sumatran orangutans were also much less affected by the Pleistocene climate changes and experienced a remarkably stable population history, while those on Borneo underwent multiple bottlenecks and a long-term decline (Chapters 4 and 5). The fact that Sumatran orangutans did not have to cope with the same environmental constrains as especially the northeastern Bornean orangutans, has likely facilitated a very different adaptive evolutionary history (Chapter 6), which is in agreement with the well-documented differences in phenotypic traits (Wich *et al.* 2009b).

The aforementioned conditions have probably allowed Sumatran orangutans to develop/maintain (the ancestral state is unknown) their larger brains (Taylor & van Schaik 2007), their higher sociability (e.g. Mitra Setia *et al.* 2009; Weingrill *et al.* 2011), and likely linked to both, their larger and more complex cultural repertoire (van Schaik 2004; van Schaik *et al.* 2009a; Krützen *et al.* 2011). In agreement with this hypothesis, we found signatures for potential positive selection within genes having crucial functions in learning, memory, and adult brain plasticity (Chapter 6). Selective changes in these genes may provide Sumatran orangutans a framework for extended flexibility (behavioral plasticity) related to both individual and social learning (van Schaik *et al.* 2003). In addition, probably linked their higher sociability, one of the identified top-candidate genes plays an important role in the oxytocin pathway (Chapter 6).

7.2 Implications for conservation and taxonomy

The results presented herein have important implications for the taxonomy and conservation management of orangutans. First of all, our findings clearly support the classification of Bornean (*P. pygmaeus*) and Sumatran orangutans (*P. abelii*) as two distinct species. The two former subspecies were elevated to different species in a taxonomic revision over a decade ago (Groves 2001), mainly based on the results of early genetic studies using mtDNA loci (Ryder & Chemnick 1993; Xu & Arnason 1996; Zhi *et al.* 1996; Warren *et al.* 2001; Zhang *et al.* 2001). However, this classification has been questioned (e.g. Muir *et al.* 1998) and they are still often perceived as "the orangutan". Muir *et al.* (1998) argued strongly against the elevation to species level by raising two main points: (i) the genetic data had been restricted to mitochondrial markers, and (ii) the genetic sampling had not been representative for the genus' distribution. By reconstructing the evolutionary history of orangutans using autosomal and sex-specific whole genome data as well as by gaining insights into their adaptive divergence, we demonstrated that Bornean and Sumatran orangutans are genetically highly differentiated.

The analysis of complete mitogenomes confirmed a deep and strict separation of Bornean and Sumatran female lineages (Chapter 5; Arora *et al.* 2010; Nater *et al.* 2011). Furthermore, the genomic MSY data indicates that Bornean and Sumatran orangutans are completely reproductively isolated for at least the last 300,000–400,000 years (Chapter 5), with substantially reduced levels of gene flow starting already about one million years ago, as

shown by the divergence of the autosomal gene pools (Chapter 4). Moreover, our analyses revealed that orangutans from both islands experienced drastically different evolutionary histories since the early Pleistocene (Chapters 4 and 5) and that they exhibit very distinct genetic local adaptations (Chapter 6). Bornean and Sumatran orangutans also show marked differences in their morphology, behavioral ecology, social organization, and life history (van Schaik *et al.* 2009b). Taken together, and comparing the split/divergence times and genetic differentiation of Bornean and Sumatran orangutans with those of the other great apes (Prado-Martinez *et al.* 2013), I conclude that they are valid species and are on the same level as for instance bonobos and chimpanzees.

The genomic and genetic data presented in this dissertation (Chapters 3–6) together with the data compiled in in-depth studies of non-invasively sampled wild orangutans throughout the genus' range (e.g. Arora *et al.* 2010; Nater *et al.* 2011; Nietlisbach *et al.* 2012; Nater *et al.* 2013; Nater *et al.* 2015), also provide the first comprehensive assessment of conservation units within orangutan species. Our genomic data confirm pronounced population structure on both islands. The geographic structure identified in the autosomal genome diversity (Chapter 4) matches the phylogeographic patterns of mitogenome haplotypes (Chapter 5), and is largely in agreement with the previously described genetic clusters from classical microsatellite markers (Chapter 3; Arora *et al.* 2010; Nater *et al.* 2013; Greminger *et al.* 2014; Nater *et al.* 2015). The populations are separated by geographical features acting as dispersal barriers for orangutans such as large rivers or mountain ridges (cf. Figure 1 in Chapter 4).

Most notable is the special genetic status of Batang Toru, the only extant orangutan population south of Lake Toba on Sumatra (Chapters 4 and 5; Nater *et al.* 2011; Nater *et al.* 2013; Nater *et al.* 2015). Batang Toru likely represents a remnant of a larger historical meta-population in central and south Sumatra (Rijksen & Meijaard 1999), from which gene flow with Borneo took place. We found that the Sumatran orangutans south and north of Lake Toba exhibit deep and long-lasting genetic separation. This boundary around Lake Toba, and not as expected the separation of the two currently recognized species, marks the deepest spit (around 3.5–4.0 Ma) in the mtDNA phylogeny of the genus *Pongo* (Chapter 5; Nater *et al.* 2011). This deep divergence of mtDNA lineages within a single species is exceptional among primates (Finstermeier *et al.* 2013). Also autosomally, Batang Toru is highly distinct from the other Sumatran populations (Chapter 4; Nater *et al.* 2013) and forms a third cluster in a PCA, where the first two clusters delineate both species (Chapter 4). In line with this strong autosomal differentiation, we inferred in a recent demographic modeling study (Nater *et al.* 2015) that gene flow between the populations north and south of Lake Toba was historically very low.

In light of these results, we suggest a taxonomic revision of *P. abelii*. In orangutans, subspecies designation is based on morphological characters (Groves 2001) and early genetic data (Warren *et al.* 2001). In Borneo, three subspecies are currently recognized. The Bornean subspecies, however, are only the result of a comparatively fairly recent radiation from a common refugium during the penultimate glacial period (130–190 ka) and *P. p. morio* is even

paraphyletic for mtDNA (Chapter 5; Nater *et al.* 2011). In contrast, Sumatran orangutans are not divided into subspecies (Groves 2001), although splits between genetic clusters are much deeper than in Borneo. I suggest to study the morphology, physiology, and life history of orangutans north and south of Lake Toba with regard to subspecies status.

With respect to a special conservation status of orangutans from south of Lake Toba, orangutans from the Batang Toru population should also be treated as a separate evolutionary significant unit (ESU) in conservation management, given the long-lasting separation and genetic uniqueness. Considering the already extremely low census size of the Batang Toru population of probably less than 400 individuals (Wich *et al.* 2008; Marshall *et al.* 2009) and the fact that most of the forest in this area is not under protection (Wich *et al.* 2011a; Wich *et al.* 2014), they should be given highest conservation status

To conclude, we identified the following conservation units for the genus *Pongo*: for Sumatran orangutans, we found two ESUs (North Toba and South Toba) and at the lower level three distinct population segments (DPS) (U.S. Fish and Wildlife Service 1996), i.e. Batang Toru, West Alas, and Northeast Alas (including Langkat and North Aceh). For Bornean orangutans, we identified five DPSs, namely Central/West Kalimantan, Sarawak, East Kalimantan, North Kinabatangan, and South Kinabatangan. Based solely on neutral genetic differentiation, Bornean orangutans form only one ESU. However, there are strong indications that the orangutans in northeastern Borneo exhibit distinct genetic adaptations to a harsher local environment with increased impact of recurrent ENSO events (detailed in Chapter 6; van Schaik *et al.* 2009b). Thus, it needs to be further investigated whether the Bornean subspecies *P. p. wurmbii* (Central/West Kalimantan) and *P. p. morio* (East Kalimantan, North and South Kinabatangan) should be elevated to separate ESUs.

The identified DPS should be managed separately and DPS membership should be taken into consideration for releases of rehabilitant orangutans into wild populations. Moreover, considering the extremely male-biased dispersal system of orangutans, it is critical to keep up/restore important male migration forest routes to maintain genetic exchange and avoid inbreeding (e.g. Hedrick & Kalinowski 2000; Reed & Frankham 2003). The pronounced geographic structure of orangutan genetic diversity considerably increases risks for losing a substantial part of the total genetic diversity with all its negative consequences through continuing habitat destruction and fragmentation. Both orangutan species are still experiencing drastic population declines (Rijksen & Meijaard 1999; Singleton *et al.* 2004; Sharma *et al.* 2012a). Therefore, urgent actions need to be taken towards orangutan conservation, in particular with regards to the population of Batang Toru.

7.3 Outlook

The results, genomic data, and technical resources presented in this dissertation lay foundation for many avenues of future research. The extensive whole-genome and reduced-

representation sequencing data of orangutans with known population provenance provides a unique resource for continuing in-depth studies of orangutan population genomics.

The main future direction will be to study genetic basis and mechanisms underlying local adaptations within the genus *Pongo* more comprehensively and in greater depth, applying statistical tests targeting different time frames and modes of selection, as well as increasing both sample size and coverage. In ongoing projects, for which we have formed additional collaborations, we are validating our findings and expanding the set of selection tests. One project in progress is for instance focused on applying advanced codon-based modeling to explore selection at older time scales. Another study is investigating the adaptive evolution of bitter-taste receptor genes in detail. Further biological pathways and candidate genes are waiting to be trapped in future studies. We are also planning to examine the identified functional SNPs potentially involved in local adaptation at finer spatial scale using SNP assays applied to an extended set of non-invasively sampled wild orangutans.

Current efforts further include the statistical phasing of the genome data and the generation of species-specific recombination maps. These data will be highly valuable in several respects. For instance, they will enable the use of haplotype-based tests (e.g. Sabeti *et al.* 2002; Voight *et al.* 2006; Chen *et al.* 2010; Pybus *et al.* in review) to detect recent hard sweeps and hopefully also soft sweep patterns. They will also facilitate exploring the genomic landscape of speciation and identifying genomic regions that show signals of adaptive introgression (Hedrick 2013). Moreover, the species-specific recombination maps will allow us to derive significance thresholds for the genome scans at which the null hypothesis of neutral evolution can be rejected through simulating genomic regions under the inferred demographic model (discussed above) and considering the local recombination rates.

Disentangling the different evolutionary processes shaping genetic diversity and unambiguously demonstrating the action of positive selection remains a challenging endeavor however (Crisci *et al.* 2012; Crisci *et al.* 2013). It is necessary to develop sophisticated methods to jointly infer demographic history, population structure, and selection simultaneously (Li *et al.* 2012; Bank *et al.* 2014). Of great importance is also to co-estimate background selection, recognizing that this process is probably prevalent (Zeng & Charlesworth 2011; Bank *et al.* 2014), as likely also true for orangutans (Chapter 6; Ma *et al.* 2013). Building up on a solid theoretical basis, the development of such novel methods is a major current focus of the field, with some promising approaches emerging (Bazin *et al.* 2010; Li *et al.* 2012; Barthelmé & Chopin 2014).

Aside from analytical advancements, for the study of non-human great apes (and many other large mammals), a key requirement for future studies on fine-scale local adaptation is to develop methods that facilitate generating genomic data from large numbers of non-invasively sampled wild individuals. This would ultimately allow taking population genetics to true population genomics. In case of orangutans, due to own efforts and extensive collaborations, the Anthropological Institute and Museum curates an extraordinary sample

collection of more than three-thousand fecal and hair samples from wild orangutans throughout the genus' range, which will represent an invaluable basis for such efforts. DNA of non-invasively collected samples is of very low quantity and quality (highly degraded), and contains only painfully low levels of endogenous DNA in the case of fecal samples. These characteristics preclude them from shotgun sequencing or common reduced genome complexity sequencing approaches as described in Chapter 3. A promising approach to sequence genomic-scale data from such samples nevertheless could be to use capture-based enrichment strategies, such as biotinylated RNA baits transcribed from genomic DNA libraries (Carpenter *et al.* 2013). This would then allow in-solution capture of the endogenous DNA in extracts from non-invasively collected samples. In fact, this strategy has already been successfully applied to ancient DNA (Carpenter *et al.* 2013; Enk *et al.* 2014). Because the RNA baits are generated in-house, this is a cost-effective way of enriching target DNA for high-throughput sequencing. To capture only parts of the genome, baits produced synthetically have proven to be highly useful (Perry *et al.* 2010). If the interest is restricted only few specific genomic regions, long-range PCR products amplified from high-quality DNA samples may be used as baits to enrich library pools (Maricic *et al.* 2010; Peñalba *et al.* 2014).

Population genomic data from non-invasive samples will give us the exciting opportunity to study the genetic basis underlying the systematic variation of orangutan phenotypic traits at the smallest local scale possible given the distribution of extant orangutan populations. Related to this, it also offers the compelling possibility to link spatially distinct environmental heterogeneity to adaptive genetic variation in modeling frameworks incorporating large environmental datasets (Manel *et al.* 2003; Schoville *et al.* 2012). The results of such studies will also be relevant for the conservation management of orangutans, by for instance clarifying if the Bornean subspecies *P. p. morio* needs to be elevated to a separate ESU, as suggested by my results.

Other interesting avenues of future research in orangutans are facilitated by the genomic MSY data I generated in the scope of this dissertation. The identified MSY-specific SNPs and microsatellite markers can be used to establish multiplex assays that allow large-scale genotyping of male-specific markers in a large number of wild orangutans. Such data would for instance provide great power for fine scale analyses of male-mediated gene flow rates among regions to estimate genetic connectedness, which is also highly relevant for conservation management. Moreover, large-scale male-specific data would allow studying potential male reproductive skew in Sumatran orangutans more comprehensively (Chapter 5; Goossens *et al.* 2006b; Dunkel *et al.* 2013; Lenzi 2014), by separating genetic signals of reproductive skew from those generated by population structure and gene flow.

Genomics has profoundly changed the fields of population genetics, molecular ecology and conservation biology. Over the past few years, we have witnessed an incredibly rapid development of high-throughput sequencing techniques, bioinformatical analysis tools and theoretical frameworks. Evolutionary genomics is starting to get out of its infancy years now, and the focus is turning from comparative research between species to population-based

studies within species. We still face, however, many challenges related to both the generation and analysis of genomic data. To avoid drawing erroneous conclusions, appropriate analysis methods need to be chosen and results need to be interpreted carefully and in a way that takes the organisms' biology into account. Undoubtedly, the coming years will bring many exciting new discoveries and improved understanding of how the different evolutionary processes shape genome-wide patterns of genetic diversity.

List of additional publications

Sonay TB, Carvalho T, Robinson M, **Greminger MP**, Krützen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Research* (in press).

Nater A, **Greminger MP**, Arora N, van Schaik CP, Goossens B, Singleton I, Verschoor EJ, Warren KS, Krützen M (2015) Reconstructing the Demographic History of Orang-utans using Approximate Bayesian Computation. *Molecular Ecology* 24, 310-327.

Nussberger B, **Greminger MP**, Grossen C, Keller LF, Wandeler P (2013) Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources* 13, 447-460.

Nater A, Arora N, **Greminger MP**, van Schaik CP, Singleton I, Wich SA, Fredriksson G, Perwitasari-Farajallah D, Pamungkas J, Krützen M (2013) Marked population structure and recent migration in the critically endangered Sumatran orangutan (*Pongo abelii*). *Journal of Heredity* 104, 2-13.

Rotheray E, Lepais O, Nater A, Krützen M, **Greminger MP**, Goulson D, Bussiere L (2012) Genetic variation and population decline of an endangered hoverfly *Blera fallax* (Diptera: *Syrphidae*). *Conservation Genetics* 13, 1283-1291.

Arora N, Van Noordwijk MA, Ackermann C, Willems EP, Nater A, **Greminger MP**, Nietlisbach P, Dunkel LP, Utami Atmoko SS, Pamungkas J, Perwitasari-Farajallah D, Van Schaik CP, Krützen M (2012) Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (*Pongo pygmaeus*). *Molecular Ecology* 21, 3352-3362.

Rotheray EL, **Greminger MP**, Nater A, Krützen M, Goulson D, Bussière L (2012) Polymorphic microsatellite loci for the endangered pine hoverfly *Blera fallax* (Diptera: *Syrphidae*). *Conservation Genetics Resources* 4, 117-120.

Nietlisbach P, Nater A, **Greminger MP**, Arora N, Krützen M (2010) A multiplex-system to target 16 male-specific and 15 autosomal genetic markers for orang-utans (genus: *Pongo*). *Conservation Genetics Resources* 2, 153-158.

Greminger MP, Schäfer MA, Nater A, Blanckenhorn WU, Krützen M (2009) Development of polymorphic microsatellite markers for the dung fly (*Sepsis cynipsea*). *Molecular Ecology Resources* 9, 1554-1556.

References

Aimi M, Bakar A (1996) Distribution and deployment of Presbytis melalophos group in Sumatera, Indonesia. *Primates* **37**, 399-409.

Akey JM (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research* **19**, 711-722.

Allan R, Lindesay J, Parker D (1996) *El Nino: Southern oscillation and climatic variability* CSIRO Publishing, Collingwood, Victoria, Australia.

Almon RR, DuBois DC, Lai W*, et al.* (2009) Gene expression analysis of hepatic roles in cause and development of diabetes in Goto-Kakizaki rats. *Journal of Endocrinology* **200**, 331-346.

Altshuler D, Pollara VJ, Cowles CR*, et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-516.

Altshuler DM, Gibbs RA, Peltonen L*, et al.* (2010a) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58.

Altshuler DM, Lander ES, Ambrogio L*, et al.* (2010b) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073.

Amaral A, Megens H-J, Kerstens H*, et al.* (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**, 374.

Amaral AJ, Ferretti L, Megens H-J*, et al.* (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6**, e14782.

Ambrose SH (1998) Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *Journal of Human Evolution* **34**, 623-651.

Andolfatto P, Przeworski M (2000) A genome-wide departure from the standard neutral model in natural populations of Drosophila. *Genetics* **156**, 257-268.

Andres O, Ronn AC, Bonhomme M*, et al.* (2008) A microarray system for Y chromosomal and mitochondrial single nucleotide polymorphism analysis in chimpanzee populations. *Molecular Ecology Resources* **8**, 529-539.

Andrews S (2012) FastQC. A quality control tool for high throughput sequence data. *[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]*

Arora N, Nater A, van Schaik CP*, et al.* (2010) Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (Pongo pygmaeus). *Proceedings of the National Academy of Sciences* **107**, 21376-21381.

Arora N, Van Noordwijk MA, Ackermann C*, et al.* (2012) Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (Pongo pygmaeus). *Molecular ecology* **21**, 3352-3362.

Auton A, Fledel-Alon A, Pfeifer S*, et al.* (2012) A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193-198.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376.

Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences* **104**, 7489-7494.

Bamshad MJ, Watkins WS, Dixon ME*, et al.* (1998) Female gene flow stratifies Hindu castes. *Nature* **395**, 651-652.

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics* **30**, 540-546.

Barkman TJ, Simpson BB (2001) Origin of high-elevation dendrochilum species (Orchidaceae) endemic to Mount Kinabalu, Sabah, Malaysia. *Systematic Botany* **26**, 658-669.

Barrett JC, Clayton DG, Concannon P*, et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41**, 703-707.

Barthelmé S, Chopin N (2014) Expectation Propagation for Likelihood-Free Inference. *Journal of the American Statistical Association* **109**, 315-333.

Baskin B, Skinner J, Sanatani S*, et al.* (2013) TMEM43 mutations associated with arrhythmogenic right ventricular cardiomyopathy in non-Newfoundland populations. *Human Genetics* **132**, 1245-1252.

Batista-Brito R, Rossignol E, Hjerling-Leffler J*, et al.* (2009) The cell-intrinsic requirement of Sox6 for cortical interneuron development. *Neuron* **63**, 466-481.

Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* **185**, 587-602.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research* **17**, 1505-1519.

Bellott DW, Hughes JF, Skaletsky H*, et al.* (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494-499.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580.

Bergstrom DE, Grieco DA, Sonti MM*, et al.* (1998) The mouse Y chromosome: Enrichment sizing, and cloning by bivariate flow cytometry. *Genomics* **48**, 304-313.

Bersaglieri T, Sabeti PC, Patterson N*, et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**, 1111 - 1120.

Beukelaers P, Vandenbosch R, Caron N*, et al.* (2011) Cdk6-Dependent Regulation of G1 Length Controls Adult Neurogenesis. *STEM CELLS* **29**, 713-724.

Bidon T, Janke A, Fain SR*, et al.* (2014) Brown and polar bear Y chromosomes reveal extensive male-biased gene flow within brother lineages. *Molecular biology and evolution*, msu109.

Bininda-Emonds ORP, Cardillo M, Jones KE*, et al.* (2007) The delayed rise of present-day mammals. *Nature* **446**, 507-512.

Bird MI, Taylor D, Hunt C (2005) Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? *Quaternary Science Reviews* **24**, 2228-2242.

Blanchette M, Bataille AR, Chen X*, et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research* **16**, 656-668.

Bowman AW, Azzalini A (2010) R package 'sm': nonparametric smoothing methods (version 2.2-4)

Boyd JL, Skove Stephanie L, Rouanet Jeremy P*, et al.* (2015) Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex. *Current Biology* **25**, 772-779.

Brandon-Jones D, Eudey AA, Geissmann T*, et al.* (2004) Asian Primate Classification. *International Journal of Primatology* **25**, 97-164.

Brown P, Sutikna T, Morwood MJ*, et al.* (2004) A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055-1061.

Brunet M, Guy F, Pilbeam D*, et al.* (2002) A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145-151.

Caballero A (1995) On the Effective Size of Populations With Separate Sexes, With Particular Reference to Sex-Linked Genes. *Genetics* **139**, 1007-1011.

Campell J (1949) *The Hero with a Thousand Faces* Princton University Press, Princeton.

Cannon CH, Manos PS (2003) Phylogeography of the Southeast Asian stone oaks (Lithocarpus). *Journal of Biogeography* **30**, 211-226.

Cannon CH, Morley RJ, Bush AB (2009) The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. *Proceedings of the National Academy of Sciences* **106**, 11188-11193.

Cao J, Schneeberger K, Ossowski S*, et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics* **43**, 956-963.

Caron N, Genin E, Vandenbosch R*, et al.* (2014) Neuronal Differentiation in the Adult Brain: CDK6 as the Molecular Regulator. In: *Tumors of the Central Nervous System, Volume 12* (ed. Hayat MA), pp. 19-32. Springer Netherlands.

Carpenter Meredith L, Buenrostro Jason D, Valdiosera C*, et al.* (2013) Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics* **93**, 852-864.

Carter CS (2014) Oxytocin Pathways and the Evolution of Human Behavior. *Annual Review of Psychology* **65**, 17-39.

Chan Y-C, Roos C, Inoue-Murayama M*, et al.* (2012) A comparative analysis of Y chromosome and mtDNA phylogenies of the Hylobates gibbons. *BMC evolutionary biology* **12**, 150.

Charlesworth B (2013) Background Selection 20 Years on The Wilhelmine E. Key 2012 Invitational Lecture. *Journal of Heredity* **104**, 161-171.

Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **355**, 1563-1572.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**, 155-174.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research* **20**, 393-402.

Chen Y-K, Chen C-Y, Hu H-T, Hsueh Y-P (2012) CTTNBP2, but not CTTNBP2NL, regulates dendritic spinogenesis and synaptic distribution of the striatin–PP2A complex. *Molecular Biology of the Cell* **23**, 4383-4392.

Chen Y-K, Hsueh Y-P (2012) Cortactin-Binding Protein 2 Modulates the Mobility of Cortactin and Regulates Dendritic Spine Formation and Maintenance. *The Journal of Neuroscience* **32**, 1043-1055.

Chesner C, Rose W, Deino A, Drake R, Westgate J (1991) Eruptive history of Earth's largest Quaternary caldera (Toba, Indonesia) clarified. *Geology* **19**, 200-203.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**, 983-995.

Cho YS, Go MJ, Kim YJ*, et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics* **41**, 527-534.

Chuang H-C, Huang T-N, Hsueh Y-P (2014) Neuronal excitation upregulates Tbr1, a high-confidence risk gene of autism, mediating Grin2b expression in the adult brain. *Frontiers in Cellular Neuroscience* **8**, 280.

Cohen K, Gibbard P (2011) Global chronostratigraphical correlation table for the last 2.7 million years. Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy), Cambridge, England.

Comuzzie AG, Cole SA, Laston SL*, et al.* (2012) Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS one* **7**, e51954.

Crisci JL, Poh Y-P, Bean A, Simkin A, Jensen JD (2012) Recent Progress in Polymorphism-Based Population Genetic Inference. *Journal of Heredity* **103**, 287-296.

Crisci JL, Poh Y-P, Mahajan S, Jensen JD (2013) The Impact of Equilibrium Assumptions on Tests of Selection. *Frontiers in Genetics* **4**.

Croteau-Chonka DC, Marvelle AF, Lange EM*, et al.* (2011) Genome-Wide Association Study of Anthropometric Traits and Evidence of Interactions With Age and Study Year in Filipino Women. *Obesity* **19**, 1019-1027.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**, 3133-3157.

Cutler DJ, Jensen JD (2010) To Pool, or Not to Pool? *Genetics* **186**, 41-43.

Danecek P, Auton A, Abecasis G*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772-772.

Darwin C, Wallace AR (1858) *Proceedings of Linnean Society of London* **3**, 45.

de Bruyn M, Stelbrink B, Morley RJ*, et al.* (2014) Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. *Systematic biology* **63**, 879-901.

de Vernal A, Hillaire-Marcel C (2008) Natural variability of Greenland climate, vegetation, and ice volume during the past million years. *Science* **320**, 1622-1625.

Delgado RA, Lameira A, Davila Ross M*, et al.* (2009) Geographical variation in orangutan long calls (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 215-224. Oxford University Press

Delgado RA, van Schaik CP (2000) The behavioral ecology and conservation of the orangutan (Pongo pygmaeus): A tale of two islands. *Evolutionary Anthropology* **9**, 201-218.

den Hoed M, Eijgelsheim M, Esko T*, et al.* (2013) Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics* **45**, 621-631.

DePristo MA, Banks E, Poplin R*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498.

Derrien T, Estellé J, Marco Sola S*, et al.* (2012) Fast Computation and Applications of Genome Mappability. *PLoS ONE* **7**, e30377.

Di Benedetto A, Sun L, Zambonin CG*, et al.* (2014) *Osteoblast regulation via ligand-activated nuclear trafficking of the oxytocin receptor*.

Dierenfeld E (1997) Orangutan nutrition. In: *Orangutan SSP husbandry manual*, Brookfield, IL: Orangutan SSP and Brookfield Zoo.

Dilley M, Heyman BN (1995) ENSO and Disaster: Droughts, Floods and El Niño/Southern Oscillation Warm Events. *Disasters* **19**, 181-193.

Doherty D, Chudley Albert E, Coghlan G*, et al.* (2012) GPSM2 Mutations Cause the Brain Malformations and Hearing Loss in Chudley-McCullough Syndrome. *The American Journal of Human Genetics* **90**, 1088-1093.

Douadi MI, Gatti S, Levrero F*, et al.* (2007a) Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). *Molecular Ecology* **16**, 2247-2259.

Douadi MI, Gatti S, Levréro F*, et al.* (2007b) Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). *Molecular Ecology* **16**, 2247-2259.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS biology* **4**, e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29**, 1969-1973.

Duforet-Frebourg N, Bazin E, Blum MG (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular biology and evolution*, msu182.

Duforet-Frebourg N, Laval G, Bazin E, Blum MG (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *arXiv preprint arXiv:1504.04543*.

Dunkel LP, Arora N, van Noordwijk MA*, et al.* (2013) Variation in developmental arrest among male orangutans: a comparison between a Sumatran and a Bornean population. *Front. Zool.*

Elderfield H, Ferretti P, Greaves M*, et al.* (2012) Evolution of Ocean Temperature and Ice Volume Through the Mid-Pleistocene Climate Transition. *Science* **337**, 704-709.

Elferink MG, Megens H-J, Vereijken A*, et al.* (2012) Signatures of selection in the genomes of commercial and non-commercial chicken breeds. *PLoS ONE* **7**, e32720.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435-445.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* **29**, 51-63.

Elshire RJ, Glaubitz JC, Sun Q*, et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379.

Enard D, Depaulis F, Roest Crollius H (2010) Human and Non-Human Primate Genomes Share Hotspots of Positive Selection. *PLoS Genet* **6**, e1000840.

Enard W (2014) Comparative genomics of brain size evolution. *Frontiers in Human Neuroscience* **8**, 345.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Enk JM, Devault AM, Kuch M*, et al.* (2014) Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. *Molecular Biology and Evolution* **31**, 1292-1294.

Eriksson J, Siedel H, Lukas D*, et al.* (2006) Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (Pan paniscus). *Molecular Ecology* **15**, 939-949.

Erler A, Stoneking M, Kayser M (2004) Development of Y-chromosomal microsatellite markers for nonhuman primates. *Molecular Ecology* **13**, 2921-2930.

Esteve-Codina A, Kofler R, Himmelbauer H*, et al.* (2011) Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity* **107**, 256-264.

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11**, 1591-1604.

Evans BJ, Supriatna J, Melnick DJ (2001) Hybridization and Population Genetics of Two Macaque Species in Sulawesi, Indonesia. *Evolution* **55**, 1686-1702.

Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* **11**, 93-108.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Faircloth BC (2008) msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources* **8**, 92-94.

Falk D, Hildebolt C, Smith K*, et al.* (2005) The brain of LB1, Homo floresiensis. *Science* **308**, 242-245.

Finstermeier K, Zinner D, Brameier M*, et al.* (2013) A mitogenomic phylogeny of living primates. *PLoS One* **8**, e69504.

Fisher RA (1930) *The genetical theory of natural selection: a complete variorum edition* Oxford University Press.

Flenley J (1998) Tropical forests under the climates of the last 30,000 years. In: *Potential Impacts of Climate Change on Tropical Forest Ecosystems*, pp. 37-57. Springer.

Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome research* **23**, 1089-1096.

Frieden LA, Townsend TA, Vaught DB*, et al.* (2010) Regulation of heart valve morphogenesis by Eph receptor ligand, ephrin-A1. *Developmental Dynamics* **239**, 3226-3234.

Gagnon J, Anini Y (2013) Glucagon Stimulates Ghrelin Secretion Through the Activation of MAPK and EPAC and Potentiates the Effect of Norepinephrine. *Endocrinology* **154**, 666-674.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**, 915-934.

Gathorne-Hardy F, Davies R, Eggleton P, Jones D (2002) Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. *Biological Journal of the Linnean Society* **75**, 453-466.

Gathorne-Hardy FJ, Harcourt-Smith WEH (2003) The super-eruption of Toba, did it cause a human bottleneck? *Journal of Human Evolution* **45**, 227-230.

Gaveau DL, Wich S, Epting J*, et al.* (2009) The future of forests and orangutans (Pongo abelii) in Sumatra: predicting impacts of oil palm plantations, road construction, and mechanisms for reducing carbon emissions from deforestation. *Environmental Research Letters* **4**, 034013.

Gibbs RA, Rogers J, Katze MG*, et al.* (2007) Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* **316**, 222-234.

Gibbs RA, Taylor JF, Van Tassell CP*, et al.* (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528-532.

Goetting-Minesky MP, Makova KD (2006) Mammalian male mutation bias: Impacts of generation time and regional variation in substitution rates. *Journal of Molecular Evolution* **63**, 537-544.

Goldstein JL, Zhao T-j, Li RL*, et al.* (2011) Surviving Starvation: Essential Role of the Ghrelin-Growth Hormone Axis. *Cold Spring Harbor Symposia on Quantitative Biology* **76**, 121-127.

Goossens B, Chikhi L, Ancrenaz M*, et al.* (2006a) Genetic Signature of Anthropogenic Population Collapse in Orang-utans. *PLoS Biology* **4**, e25.

Goossens B, Chikhi L, Jalil M*, et al.* (2005) Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (Pongo pygmaeus) populations from Sabah, Malaysia. *Molecular Ecology* **14**, 441-456.

Goossens B, Setchell J, James S*, et al.* (2006b) Philopatry and reproductive success in Bornean orang-utans (Pongo pygmaeus). *Molecular Ecology* **15**, 2577-2588.

Gorog AJ, Sinaga MH, Engstrom MD (2004) Vicariance or dispersal? Historical biogeography of three Sunda shelf murine rodents (Maxomys surifer, Leopoldamys sabanus and Maxomys whiteheadi). *Biological Journal of the Linnean Society* **81**, 91-109.

Goudet J, Perrin N, Waser P (2002) Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Molecular Ecology* **11**, 1103-1114.

Graves JAM, Wakefield MJ, Toder R (1998) The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human Molecular Genetics* **7**, 1991-1996.

Greenwood PJ (1980) Mating Systems, Philopatry and Dispersal in Birds and Mammals. *Animal Behaviour* **28**, 1140-1162.

Greminger M (2007) *The quest for the Y - Development and application of male-specific markers in orangutans (Pongo sp.) and bottlenose dolphins (Tursiops sp.)* Master's thesis, University of Zürich, Switzerland.

Greminger MP, Kruetzen M, Schelling C, Pienkowska–Schelling A, Wandeler P (2010) The quest for Y-chromosomal markers–methodological strategies for mammalian non-model organisms. *Molecular ecology resources* **10**, 409-420.

Greminger MP, Stolting K, Nater A*, et al.* (2014) Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC genomics* **15**, 16.

Grossman Sharon R, Andersen Kristian G, Shlyakhter I*, et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703-713.

Groves C (2001) *Primate Taxonomy* Smithsonian Books.

Gusmao L, Sanchez-Diz P, Calafell F*, et al.* (2005) Mutation rates at Y chromosome specific microsatellites. *Human Mutation* **26**, 520-528.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**, 790-799.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution* **62**, 255-265.

Hailer F, Leonard JA (2008) Hybridization among three native North American Canis species in a region of natural sympatry. *Plos one* **3**, e3333.

Hall R (1996) Reconstructing Cenozoic SE Asia. In: *Tectonic Evolution of Southeast Asia*, pp. 153-184. The Geological Society, London.

Hall R (2002) Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. *Journal of Asian Earth Sciences* **20**, 353-431.

Hall R (2013) The palaeogeography of Sundaland and Wallacea since the Late Jurassic. *Journal of Limnology* **72**, e1.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT **41**, 95-98.

Halligan DL, Kousathanas A, Ness RW*, et al.* (2013) Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. *PLoS Genet* **9**, e1003995.

Hammond RL, Lawson Handley LJ, Winney BJ, Bruford MW, Perrin N (2006) Genetic evidence for female-biased dispersal and gene flow in a polygynous primate. *Proceedings of the Royal Society B: Biological Sciences* **273**, 479-484.

Handley L, Perrin N (2007a) Advances in our understanding of mammalian sex-biased dispersal. *Molecular Ecology* **16**, 1559-1578.

Handley LJL, Berset-Brandli L, Perrin N (2006a) Disentangling reasons for low Y chromosome variation in the greater white-toothed shrew (Crocidura russula). *Genetics* **173**, 935-942.

Handley LJL, Hammond RL, Emaresi G, Reber A, Perrin N (2006b) Low Y chromosome variation in Saudi-Arabian hamadryas baboons (Papio hamadryas hamadryas). *Heredity* **96**, 298-303.

Handley LJL, Perrin N (2006) Y chromosome microsatellite isolation from BAC clones in the greater white-toothed shrew (Crocidura russula). *Molecular Ecology Notes* **6**, 276-279.

Handley LLJ, Perrin N (2007b) Advances in our understanding of mammalian sex-biased dispersal. *Molecular Ecology* **16**, 1559-1578.

Hanemaaijer N, Dijkhuizen T, Haadsma M*, et al.* (2009) A 649 kb microduplication in 1p34.1, including POMGNT1, in a patient with microcephaly, coloboma and laryngomalacia; and a review of the literature. *European Journal of Medical Genetics* **52**, 116-119.

Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths.

Harrison T, Krigbaum J, Manser J (2006) Primate biogeography and ecology on the Sunda Shelf islands: A paleontological and zooarchaeological perspective. In: *Primate Biogeography* (eds. Lehman S, Fleagle J), pp. 331–374. Springer, New York.

Haslam M, Petraglia M (2010) Comment on "Environmental impact of the ~73kya Toba super-eruption in South Asia" by M.A.J. Williams, S.H. Ambrose, S. van der Kaars, C. Ruehlemann, U. Chattopadhyaya, J. Pal and P.R. Chauhan [Palaeogeography, Palaeoclimatology, Palaeoecology 284 (2009) 295–314]. *Palaeogeography, Palaeoclimatology, Palaeoecology* **296**, 199-203.

Head MJ, Gibbard PL (2005) Early-Middle Pleistocene transitions: the land-ocean evidence.

Hébert JM, McConnell SK (2000) Targeting of cre to the Foxg1 (BF-1) Locus Mediates loxP Recombination in the Telencephalon and Other Developing Head Structures. *Developmental Biology* **222**, 296-306.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology* **22**, 4606-4618.

Hedrick PW, Kalinowski ST (2000) Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics*, 139-162.

Heinrichs M, von Dawans B, Domes G (2009) Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology* **30**, 548-557.

Hellborg L, Ellegren H (2003) Y chromosome conserved anchored tagged sequences (YCATS) for the analysis of mammalian male-specific DNA. *Molecular Ecology* **12**, 283-291.

Hellborg L, Ellegren H (2004) Low levels of nucleotide diversity in mammalian Y chromosomes. *Molecular Biology and Evolution* **21**, 158-163.

Helyar SJ, Hemmer-Hansen J, Bekkevold D*, et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 123-136.

Henneberg M, Eckhardt RB, Chavanaves S, Hsü KJ (2014) Evolved developmental homeostasis disturbed in LB1 from Flores, Indonesia, denotes Down syndrome and not diagnostic traits of the invalid species Homo floresiensis. *Proceedings of the National Academy of Sciences* **111**, 11967-11972.

Herculano-Houzel S (2012) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences* **109**, 10661-10668.

Hernandez RD, Kelley JL, Elyashiv E*, et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920-924.

Heyer E, Chaix R, Pavard S, Austerlitz F (2012) Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology* **21**, 597-612.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics* **6**, 799-803.

Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**, e1000862.

Huang T-N, Chuang H-C, Chou W-H*, et al.* (2014) Tbr1 haploinsufficiency impairs amygdalar axonal projections and results in cognitive abnormality. *Nat Neurosci* **17**, 240-247.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Notes* **8**, 3-17.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 44.

Hughes JF, Rozen S (2012) Genomics and genetics of human and primate Y chromosomes. *Annual review of genomics and human genetics* **13**, 83-108.

Hughes JF, Skaletsky H, Brown LG*, et al.* (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82-86.

Hughes JF, Skaletsky H, Pyntikova T*, et al.* (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536-539.

Hughes JF, Skaletsky H, Pyntikova T*, et al.* (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**, 100-103.

Husson SJ, Wich SA, Marshall AJ*, et al.* (2009) Orangutan distribution, density, abundance and impacts of disturbance. In: *Orangutans: Geographic variation in behavioral ecology and conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 77-96.

Hvilsom C, Qian Y, Bataillon T*, et al.* (2012) Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences* **109**, 2054-2059.

Hyten D, Cannon S, Song Q*, et al.* (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38.

Ibrahim YK, Tshen LT, Westaway KE*, et al.* (2013) First discovery of Pleistocene orangutan (Pongo sp.) fossils in Peninsular Malaysia: Biogeographic and paleoenvironmental implications. *Journal of Human Evolution* **65**, 770-797.

Isler K (2014) Adipose Tissue in Evolution. In: *Adipose Tissue and Adipokines in Health and Disease* (eds. Fantuzzi G, Braunschweig C), pp. 3-13. Humana Press.

Isler K, van Schaik CP (2009) The expensive brain: a framework for explaining evolutionary changes in brain size. *Journal of Human Evolution* **57**, 392-400.

IUCN (2014) The IUCN Red List of Threatened Species, Version 2014.3, http://www.iucnredlist.org, last accessed December 14, 2014.

Jablonski NG (1998) The response of catarrhine primates to Pleistocene environmental fluctuations in East Asia. *Primates* **39**, 29-37.

Jalil MF, Cable J, Sinyor J*, et al.* (2008) Riverine effects on mitochondrial structure of Bornean orang-utans (Pongo pygmaeus) at two spatial scales. *Molecular ecology* **17**, 2898-2909.

Jensen JD (2014) On the unfounded enthusiasm for soft selective sweeps. *Nature communications* **5**.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics* **170**, 1401-1410.

Jiang ZW, Zhang XQ, Deka R, Jin L (2005) Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Research* **33**.

Jones C (2011) Prospects of recovering ancient DNA from Homo floresiensis boosted by study on teeth. *Nature*, published online 5 January

Jonker RM, Zhang Q, Van Hooft P*, et al.* (2012) The development of a genome wide SNP set for the Barnacle Goose (Branta leucopsis). *PLoS ONE* **7**, e38412.

Kanamori T, Kuze N, Bernard H, Malim TP, Kohshima S (2010) Feeding ecology of Bornean orangutans (Pongo pygmaeus morio) in Danum Valley, Sabah, Malaysia: a 3-year record including two mast fruitings. *American Journal of Primatology* **72**, 820-840.

Kanthaswamy S, Kurushima JD, Smith DG (2006) Inferring Pongo conservation units: a perspective based on microsatellite and mitochondrial DNA analyses. *Primates* **47**, 310-321.

Kanthaswamy S, Smith DG (2002) Population subdivision and gene flow among wild orangutans. *Primates* **43**, 315-327.

Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. *Database* **2011**, bar049.

Kawamoto Y, Kawamoto S, Matsubayashi K*, et al.* (2008a) Genetic diversity of longtail macaques (Macaca fascicularis) on the island of Mauritius: an assessment of nuclear and mitochondrial DNA polymorphisms. *Journal of Medical Primatology* **37**, 45-54.

Kawamoto Y, Tomari KI, Kawai S, Kawamoto S (2008b) Genetics of the Shimokita macaque population suggest an ancient bottleneck. *Primates* **49**, 32-40.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* **7**, 1225-1241.

Kayser M, Kittler R, Erler A*, et al.* (2004) A comprehensive survey of human Y-chromosomal microsatellites. *American Journal of Human Genetics* **74**, 1183-1197.

Kayser M, Sajantila A (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Science International* **118**, 116-121.

Ke Y, Su B, Song X*, et al.* (2001) African Origin of Modern Humans in East Asia: A Tale of 12,000 Y Chromosomes. *Science* **292**, 1151-1153.

Keifer J, Zheng Z (2010) AMPA receptor trafficking and learning. *European Journal of Neuroscience* **32**, 269-277.

Kerstens H, Crooijmans R, Dibbits B*, et al.* (2011) Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries. *BMC Genomics* **12**, 94.

Kerstens H, Crooijmans R, Veenendaal A*, et al.* (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**, 479.

Kimura M (1984) *The neutral theory of molecular evolution* Cambridge University Press.

Kirsch S, Mannch C, Jiang Z*, et al.* (2008) Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions. *Genome Research* **18**, 1030-1042.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK (2011) Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC ecology* **11**, 10.

Knott C, Beaudrot L, Snaith T*, et al.* (2008) Female-Female Competition in Bornean Orangutans. *International Journal of Primatology* **29**, 975-997.

Knott CD (1998) Changes in Orangutan Caloric Intake, Energy Balance, and Ketones in Response to Fluctuating Fruit Availability. *International Journal of Primatology* **19**, 1061-1079.

Kofler R, Schlötterer C (2012) Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* **28**, 2084-2085.

Konczal M, Babik W, Radwan J, Sadowska ET, Koteja P (2015) Initial molecular-level response to artificial selection for increased aerobic metabolism occurs primarily via changes in gene expression. *Molecular Biology and Evolution*.

Kortüm F, Das S, Flindt M*, et al.* (2011) The core FOXG1 syndrome phenotype consists of postnatal microcephaly, severe mental retardation, absent language, dyskinesia, and corpus callosum hypogenesis. *Journal of Medical Genetics*.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. *Nature* **435**, 673-676.

Kosiol C, Vinař T, da Fonseca RR*, et al.* (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet* **4**, e1000144.

Kraus R, Kerstens H, Van Hooft P*, et al.* (2011) Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). *BMC Genomics* **12**, 150.

Krützen M, Willems EP, van Schaik CP (2011) Culture and geographic variation in orangutan behavior. *Current Biology* **21**, 1808-1812.

Kumar S, Schiffer PH, Blaxter M (2012) 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects. *Nucleic acids research* **40**, D1295-D1300.

Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* **286**, 964-967.

Langergraber KE, Siedel H, Mitani JC*, et al.* (2007) The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. *PLoS One* **2**, e973.

Lango Allen H, Estrada K, Lettre G*, et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838.

Laughlin SB, van Steveninck RRdR, Anderson JC (1998) The metabolic cost of neural information. *Nature Neuroscience* **1**, 36-41.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS genetics* **8**, e1002453.

Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research* **21**, 952-960.

Lee KE, Seo J, Shin J*, et al.* (2014) Positive feedback loop between Sox2 and Sox6 inhibits neuronal differentiation in the developing central nervous system. *Proceedings of the National Academy of Sciences* **111**, 2794-2799.

Leighton M (1993) Modeling dietary selectivity by Bornean orangutans: evidence for integration of multiple criteria in fruit selection. *International Journal of Primatology* **14**, 257-313.

Lent R, Azevedo FAC, Andrade-Moraes CH, Pinto AVO (2012) How many neurons do you have? Some dogmas of quantitative neuroscience under revision. *European Journal of Neuroscience* **35**, 1-9.

Lenzi I (2014) *Reproductive success and paternity concentration in wild male orangutans (genus Pongo)* MSc thesis, University of Zurich.

Leuner B, Caponiti JM, Gould E (2012) Oxytocin stimulates adult neurogenesis even under conditions of stress and elevated glucocorticoids. *Hippocampus* **22**, 861-868.

Lewontin R, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175-195.

Lewontin RC (1974) *The genetic basis of evolutionary change* Columbia University Press New York.

Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496.

Li H, Handsaker B, Wysoker A*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li J, Li H, Jakobsson M*, et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology* **21**, 28-44.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.

Locke DP, Hillier LW, Warren WC*, et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology* **23**, 2178-2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* **24**, 1031-1046.

Lugon-Moulin N, Hausser J (2002) Phylogeographical structure, postglacial recolonization and barriers to gene flow in the distinctive Valais chromosome race of the common shrew (Sorex araneus). *Molecular Ecology* **11**, 785-794.

Luo S-J, Johnson WE, David VA*, et al.* (2007) Development of Y Chromosome Intraspecific Polymorphic Markers in the Felidae. *Journal of Heredity* **98**, 400-413.

Lyons LA, Laughlin TF, Copeland NG*, et al.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics* **15**, 47-56.

Ma X, Kelley JL, Eilertson K*, et al.* (2013) Population Genomic Analysis Reveals a Rich Speciation and Demographic History of Orang-utans (Pongo pygmaeus and Pongo abelii). *PloS one* **8**, e77175.

MacDonald AJ, Sankovic N, Sarre SD*, et al.* (2006) Y chromosome microsatellite markers identified from the tammar wallaby (Macropus eugenii) and their amplification in three other macropod species. *Molecular Ecology Notes* **6**, 1202-1204.

MacKinnon K, Hatta G, Halim H*, et al.* (1996) The island of Borneo. In: *The Ecology of Kalimantan*, pp. 9-68. Periplus Editions (HK) Ltd., Singapore.

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH (2011) Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS genetics* **7**, e1001319.

Mailund T, Halager AE, Westergaard M*, et al.* (2012) A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS genetics* **8**, e1003125.

Makova KD, Li W-H (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624-626.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution* **18**, 189-197.

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004.

Marshall AJ, Ancrenaz M, Brearley FQ*, et al.* (2009) The effects of forest phenology and floristics on populations of Bornean and Sumatran orangutans. In: *Orangutans. Geographic variation in behavioral ecology and conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 311–326. Oxford University Press.

Martinson DG, Pisias NG, Hays JD*, et al.* (1987) Age dating and the orbital theory of the ice ages: development of a high-resolution 0 to 300,000-year chronostratigraphy. *Quaternary research* **27**, 1-29.

Martynoga B, Morrison H, Price DJ, Mason JO (2005) Foxg1 is required for specification of ventral telencephalon and region-specific regulation of dorsal telencephalic precursor proliferation and apoptosis. *Developmental Biology* **283**, 113-127.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23-35.

Mayr E (1982) *The growth of biological thought: diversity, evolution, and inheritance* Harvard University Press.

McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303.

McLaren W, Pritchard B, Rios D*, et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070.

McManus KF, Kelley JL, Song S*, et al.* (2015) Inference of Gorilla Demographic and Selective History from Whole-Genome Sequence Data. *Molecular Biology and Evolution* **32**, 600-612.

Meijaard E (2004) *Solving mammalian riddles: a reconstruction of the Tertiary and Quaternary distribution of mammals and their palaeoenvironments in island South-East Asia* PhD thesis, Australian National University.

Meijaard E, Buchori D, Hadiprakarsa Y*, et al.* (2011) Quantifying Killing of Orangutans and Human-Orangutan Conflict in Kalimantan, Indonesia. *PLoS ONE* **6**, e27491.

Merner ND, Hodgkinson KA, Haywood AFM*, et al.* (2008) Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 Is a Fully Penetrant, Lethal Arrhythmic Disorder Caused by a Missense Mutation in the TMEM43 Gene. *The American Journal of Human Genetics* **82**, 809-821.

Mikkelsen TS, Wakefield MJ, Aken B*, et al.* (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-177.

Mita S, Thuillet A-C, Gay L*, et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22**, 1383-1399.

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. *Current Opinion in Genetics & Development* **6**, 737-742.

Mitra Setia T, Delgado R, Utami Atmoko S, Singleton I, Van Schaik C (2009) Social organization and male-female relationships. In: *Orangutans, Geographic Variations in Behavioral Ecology and Conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 245-253.

Mitra Setia T, Van Schaik CP (2007) The response of adult orang-utans to flanged male long calls: inferences about their function. *Folia Primatologica* **78**, 215-226.

Mitsushima D, Ishihara K, Sano A, Kessels HW, Takahashi T (2011) Contextual learning requires synaptic AMPA receptor delivery in the hippocampus. *Proceedings of the National Academy of Sciences* **108**, 12503-12508.

Morley RJ (2000) *Origin and evolution of tropical rain forests* John Wiley & Sons.

Morrogh-Bernard HC, Husson SJ, Knott CD*, et al.* (2009) Orangutan activity budgets and diet. In: *Orangutans. Geographic variation in behavioral ecology and conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 119-134. Oxford University Press.

Morrogh-Bernard HC, Morf NV, Chivers DJ, Krützen M (2011) Dispersal patterns of orang-utans (Pongo spp.) in a Bornean peat-swamp forest. *International Journal of Primatology* **32**, 362-376.

Muir CC, Galdikas BM, Beckenbach AT (1998) Is there sufficient evidence to elevate the orangutan of Borneo and Sumatra to separate species? *Journal of molecular evolution* **46**, 378-379.

Muir CC, Galdikas BMF, Beckenbach AT (2000) mtDNA sequence diversity of orangutans from the islands of Borneo and Sumatra. *Journal of Molecular Evolution* **51**, 471-480.

Murphy WJ, Wilkerson AJP, Raudsepp T*, et al.* (2006) Novel Gene Acquisition on Carnivore Y Chromosomes. *PLoS Genet* **2**, e43.

Musiani M, Leonard JA, Cluff HD*, et al.* (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. *Molecular Ecology* **16**, 4149-4170.

Nabholz B, Glemin S, Galtier N (2008) Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. *Molecular biology and evolution* **25**, 120-130.

Natanaelsson C, Oskarsson MCR, Angleby H*, et al.* (2006) Dog Y chromosomal DNA sequence: identification, sequencing and SNP discovery. *BMC Genetics* **7**, 6.

Nater A (2012) *Processes underlying genetic differentiation and speciation in orangutans (Pongo spp.)* Dissertation, University of Zurich.

Nater A, Arora N, Greminger MP*, et al.* (2013) Marked population structure and recent migration in the critically endangered Sumatran orangutan (Pongo abelii). *Journal of Heredity* **104**, 2-13.

Nater A, Greminger MP, Arora N*, et al.* (2015) Reconstructing the Demographic History of Orang-utans using Approximate Bayesian Computation. *Molecular Ecology* **24**, 310-327.

Nater A, Nietlisbach P, Arora N*, et al.* (2011) Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant orangutans (genus: Pongo). *Molecular Biology and Evolution* **28**, 2275-2288.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**, 3321-3323.

Nei M (1987) *Molecular evolutionary genetics* Columbia University Press.

Nei M, Suzuki Y, Nozawa M (2010) The neutral theory of molecular evolution in the genomic era. *Annual review of genomics and human genetics* **11**, 265-289.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Nielsen R, Bustamante C, Clark AG*, et al.* (2005a) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**, e170.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-451.

Nielsen R, Williamson S, Kim Y*, et al.* (2005b) Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566-1575.

Nietlisbach P, Arora N, Nater A*, et al.* (2012) Heavily male-biased long-distance dispersal of orang-utans (genus: Pongo), as revealed by Y-chromosomal and mitochondrial genetic markers. *Molecular Ecology* **21**, 3173-3186.

Nietlisbach P, Nater A, Greminger M, Arora N, Krützen M (2010) A multiplex-system to target 16 male-specific and 15 autosomal genetic markers for orang-utans (genus: Pongo). *Conservation Genetics Resources* **2**, 153-158.

Nipperess DA (2015) A separate creation: diversity, distinctiveness and conservation of Australian wildlife. In: *Austral Ark - The State of Wildlife in Australia* (eds. Stow A, Maclean N, Holwell G), pp. 1-24. Cambridge University Press.

O'Reilly PF, Birney E, Balding DJ (2008) Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Research* **18**, 1304-1313.

Olson-Manning CF, Wagner MR, Mitchell-Olds T (2012) Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews Genetics* **13**, 867-877.

Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, 935-949.

Pabinger S, Dander A, Fischer M*, et al.* (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*.

Pardo-Diaz C, Salazar C, Jiggins CD (2014) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*.

Parker JD, Rabinovitch PS, Burmer GC (1991) Targeted gene walking polymerase chain reaction. *Nucl. Acids Res.* **19**, 3055-3060.

Patel JN, Coppack SW, Goldstein DS, Miles JM, Eisenhofer G (2002) Norepinephrine Spillover from Human Adipose Tissue before and after a 72-Hour Fast. *The Journal of Clinical Endocrinology & Metabolism* **87**, 3373-3377.

Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* **185**, 907-922.

Peñalba JV, Smith LL, Tonione MA*, et al.* (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular ecology resources* **14**, 1000-1010.

Perry GH (2014) The promise and practicality of population genomics research with endangered species. *International Journal of Primatology* **35**, 55-70.

Perry GH, Marioni JC, Melsted P, Gilad Y (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular ecology* **19**, 5332-5344.

Peter BM, Wegmann D, Excoffier L (2010) Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular ecology* **19**, 4648-4660.

Petit E, Balloux F, Excoffier L (2002) Mammalian population genetics: why not Y? *Trends in Ecology & Evolution* **17**, 28-33.

Petraglia M, Korisettar R, Boivin N*, et al.* (2007) Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. *Science* **317**, 114-116.

Philander SGH (1983) El Nino Southern Oscillation phenomena. *Nature* **302**, 295-301.

Pickrell JK, Coop G, Novembre J*, et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826-837.

Pienkowska-Schelling A, Zawada M, Schelling C (2005) A canine X chromosome painting probe applied to four canid species: close relationship of a heterochromatic-like sequence between the dog and the blue fox. *Journal of Animal Breeding and Genetics* **122**, 54-59.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research* **20**, 291-300.

Posada D (2003) Using MODELTEST and PAUP* to Select a Model of Nucleotide Substitution. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Poznik GD, Henn BM, Yee M-C*, et al.* (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562-565.

Prado-Martinez J, Sudmant PH, Kidd JM*, et al.* (2013) Great ape genetic diversity and population history. *Nature* **499**, 471-475.

Primmer CR, Moller AP, Ellegren H (1996) A wide-range survey of cross-species microsatellite amplification in birds. *Molecular Ecology* **5**, 365-378.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**, R208-R215.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* **16**, 1791-1798.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Prufer K, Munch K, Hellmann I*, et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* **advance online publication**.

Prugnolle F, de Meeus T (2002) Inferring sex-biased dispersal from population genetic tools: a review. *Heredity* **88**, 161-165.

Przeworski M (2002) The signature of positive selection at randomly chosen Loci. *Genetics* **160**, 1179-1189.

Pybus M, Luisi P, Dall'Olio G*, et al.* (in review) A Machine-Learning Framework to Detect and Classify Hard Selective Sweeps in Human Population.

Quek S, Davies S, Ashton P, Itino T, Pierce N (2007) The geography of diversification in mutualistic ants: a gene's-eye view into the Neogene history of Sundaland rain forests. *Molecular ecology* **16**, 2045-2062.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

Raaum RL, Sterner KN, Noviello CM, Stewart C-B, Disotell TR (2005) Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *Journal of Human Evolution* **48**, 237-257.

Rambaut A (2012) FigTree version 1.4. *[http://tree.bio.ed.ac.uk/software/figtree/]*

Rambaut A, Suchard M, Drummond A (2013) Tracer, version 1.6. *[http://tree.bio.ed.ac.uk/software/tracer/]*

Rampias TN, Fragoulis EG, Sideris DC (2009) A hybrid-specific, polymerase chain reaction-based amplification approach for chromosomal walking. *Analytical Biochemistry* **388**, 342-344.

Rampino M, Ambrose SH (2000) Volcanic winter in the Garden of Eden: the Toba supereruption and the late Pleistocene human population crash. In: *Volcanic hazards and disasters in human antiquity*, p. 71. Geological Society of America, Boulder.

Reddy PS, Mahanty S, Kaul T*, et al.* (2008) A high-throughput genome-walking method and its use for cloning unknown flanking sequences. *Analytical Biochemistry* **381**, 248-253.

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conservation biology* **17**, 230-237.

Reimand J, Arak T, Vilo J (2011) g: Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic acids research* **39**, W307-W315.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research* **35**, W193-W200.

Rietveld CA, Esko T, Davies G*, et al.* (2014) Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences* **111**, 13790-13794.

Rijksen HD, Meijaard E (1999) *Our vanishing relative: the status of wild orang-utans at the close of the twentieth century* Kluwer Academic Publishers Dordrecht.

Roach JC, Glusman G, Smit AFA*, et al.* (2010) Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**, 636-639.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology* **21**, 2852-2862.

Rolfe D, Brown GC (1997) Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews* **77**, 731-758.

Roos C, Zinner D, Kubatko L*, et al.* (2011) Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evolutionary Biology* **11**, 77.

Rosenthal A, Stephen D, Jones C (1990) Genomic walking and sequencing by oligo-cassette mediated polymerase chain reaction. *Nucleic Acids Research* **18**, 3095-3096.

Rozen S, Skaletsky H (1999) Primer3 on the WWW for General Users and for Biologist Programmers. In: *Bioinformatics Methods and Protocols* (eds. Misener S, Krawetz S), pp. 365-386. Humana Press.

Ruas M, Gregory F, Jones R*, et al.* (2007) CDK4 and CDK6 Delay Senescence by Kinase-Dependent and p16INK4a-Independent Mechanisms. *Molecular and Cellular Biology* **27**, 4273-4282.

Ryder OA, Chemnick L (1993) Chromosomal and mitochondrial DNA variation in orang utans. *Journal of Heredity* **84**, 405-409.

Sabeti PC, Reich DE, Higgins JM*, et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837.

Sabeti PC, Schaffner SF, Fry B*, et al.* (2006) Positive natural selection in the human lineage. *Science* **312**, 1614-1620.

Safran M, Dalah I, Alexander J*, et al.* (2010) GeneCards Version 3: the human gene integrator. *Database* **2010**, baq020.

Sanchez C, Smith T, Wiedmann R*, et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10**, 559.

Sankovic N, Delbridge ML, Grutzner F*, et al.* (2006) Construction of a highly enriched marsupial Y chromosome-specific BAC sub-library using isolated Y chromosomes. *Chromosome Research* **14**, 657-664.

Saredi S, Ardissone A, Ruggieri A*, et al.* (2012) Novel POMGNT1 point mutations and intragenic rearrangements associated with muscle-eye-brain disease. *Journal of the Neurological Sciences* **318**, 45-50.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**, 807-820.

Scally A, Dutheil JY, Hillier LW*, et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175.

Scally A, Yngvadottir B, Xue Y*, et al.* (2013) A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PloS one* **8**, e65066.

Schaffner SF, Foo C, Gabriel S*, et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**, 1576-1583.

Schoville SD, Bonin A, François O*, et al.* (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics* **43**, 23-43.

Schubert M, Jónsson H, Chang D*, et al.* (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences* **111**, E5661-E5669.

Schulz H, Emeis K-C, Erlenkeuser H, von Rad U, Rolf C (2002) The Toba volcanic event and interstadial/stadial climates at the marine isotopic stage 5 to 4 transition in the northern Indian Ocean. *Quaternary Research* **57**, 22-31.

Seeb JE, Carvalho G, Hauser L*, et al.* (2011a) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* **11**, 1-8.

Seeb LW, Templin WD, Sato S*, et al.* (2011b) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources* **11**, 195-217.

Seehausen O, Butlin RK, Keller I*, et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics* **15**, 176-192.

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**, 278-280.

Senn H, Ogden R, Cezard T*, et al.* (2013) Reference-free SNP discovery for the Eurasian beaver from restriction site–associated DNA paired-end data. *Molecular ecology*.

Sharma R, Arora N, Goossens B*, et al.* (2012a) Effective Population Size Dynamics and the Demographic Collapse of Bornean Orang-Utans. *PLoS ONE* **7**, e49429.

Sharma R, Goossens B, Kun-Rodrigues C*, et al.* (2012b) Two different high throughput sequencing approaches identify thousands of de novo genomic markers for the genetically depleted Bornean elephant. *PLoS ONE* **7**, e49533.

Shendure J, Ji HL (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145.

Singleton I, Knott CD, Morrogh-Bernard HC, Wich S, Van Schaik C (2009) Ranging behavior of orangutan females and social organization. In: *Orangutans: Geographic variation in behavioral ecology and conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 205-213. Oxford University Press

Singleton I, van Schaik CP (2001) Orangutan home range size and its determinants in a Sumatran swamp forest. *International Journal of Primatology* **22**, 877-911.

Singleton I, van Schaik CP (2002) The social organisation of a population of Sumatran orang-utans. *Folia Primatologica* **73**, 1-20.

Singleton I, Wich S, Husson S*, et al.* (2004) Orangutan population and habitat viability assessment: final report. *IUCN/SSC Conservation Breeding Specialist Group, Apple Valley, MN*.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ*, et al.* (2003a) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ*, et al.* (2003b) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-U822.

Slik JF, Aiba S-I, Bastian M*, et al.* (2011) Soils on exposed Sunda Shelf shaped biogeographic patterns in the equatorial forests of Southeast Asia. *Proceedings of the National Academy of Sciences* **108**, 12343-12347.

Soh YS, Alföldi J, Pyntikova T*, et al.* (2014) Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. *Cell* **159**, 800-813.

Sotheeswaran S, Pasupathy V (1993) Distribution of resveratrol oligomers in plants. *Phytochemistry* **32**, 1083-1092.

Spencer CC, Deloukas P, Hunt S*, et al.* (2006) The influence of recombination on human genetic diversity.

Spillmann B, Dunkel LP, Van Noordwijk MA*, et al.* (2010) Acoustic Properties of Long Calls Given by Flanged Male Orang-Utans (Pongo pygmaeus wurmbii) Reflect Both Individual Identity and Context. *Ethology* **116**, 385-395.

Stadler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**, 205-216.

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution* **22**, 63-73.

Stapley J, Reger J, Feulner PGD*, et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution* **25**, 705-712.

Steinemann M, Steinemann S (1992) Degenerating Y chromosome of Drosophila miranda: a trap for retrotransposons. *Proceedings of the National Academy of Sciences* **89**, 7591-7595.

Steiner CC, Putnam AS, Hoeck PEA, Ryder OA (2013) Conservation genomics of threatened animal species. *Annu. Rev. Anim. Biosci*, 18.11–18.21.

Steiper ME (2006) Population history, biogeography, and taxonomy of orangutans (Genus: Pongo) based on a population genetic meta-analysis of multiple loci. *Journal of Human Evolution* **50**, 509-522.

Steiper ME, Young NM (2006) Primate molecular divergence dates. *Molecular phylogenetics and evolution* **41**, 384-394.

Stephen LJ, Fawkes AL, Verhoeve A, Lemke G, Brown A (2007) A critical role for the EphA3 receptor tyrosine kinase in heart development. *Developmental Biology* **302**, 66-79.

Stölting KN, Nipper R, Lindtke D*, et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology* **22**, 842-855.

Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 43-48.

Storz JF (2005) INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**, 671-688.

Sun X, Li X, Luo Y, Chen X (2000) The vegetation and climate at the last glaciation on the emerged continental shelf of the South China Sea. *Palaeogeography, palaeoclimatology, palaeoecology* **160**, 301-316.

Sundqvist AK, Bjornerfeldt S, Leonard JA*, et al.* (2006) Unequal Contribution of Sexes in the Origin of Dog Breeds. *Genetics* **172**, 1121-1128.

Sundqvist AK, Ellegren H, Olivier M, Vila C (2001) Y chromosome haplotyping in Scandinavian wolves (Canis lupus) based on microsatellite markers. *Molecular Ecology* **10**, 1959-1966.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Tajima F (1993) Measurement of DNA polymorphism. *Mechanisms of molecular evolution*, 37-59.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512-526.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular biology and evolution* **30**, 2725-2729.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences **17**, 57-86.

Taverna E, Götz M, Huttner WB (2014) The Cell Biology of Neurogenesis: Toward an Understanding of the Development and Evolution of the Neocortex. *Annual Review of Cell and Developmental Biology* **30**, 465-502.

Taylor AB (2006) Feeding behavior, diet, and the functional consequences of jaw form in orangutans, with implications for the evolution of Pongo. *Journal of Human Evolution* **50**, 377-393.

Taylor AB (2009) The functional significance of variation in jaw form in orangutans. In: *Orangutans - Geographic variation in Behavioral Ecology and Conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP). Oxford University Press.

Taylor AB, van Schaik CP (2007) Variation in brain size and ecology in Pongo. *Journal of Human Evolution* **52**, 59-71.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome research* **16**, 702-712.

Thalmann O, Serre D, Hofreiter M*, et al.* (2005) Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. *Molecular Ecology* **14**, 179-188.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073.

The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Research* **43**, D204-D212.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678.

Thinh VN, Mootnick AR, Geissmann T*, et al.* (2010) Mitochondrial evidence for multiple radiations in the evolutionary history of small apes. *BMC evolutionary biology* **10**, 74.

Thornton K, Andolfatto P (2006) Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics* **172**, 1607-1619.

Triglia T, Peterson MG, Kemp DJ (1988) A procedure for in vitro amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Research* **16**, 8186-.

Truong HT, Ramos AM, Yalcin F*, et al.* (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE* **7**, e37565.

Tsuchiya T, Kameya N, Nakamura I (2009) Straight Walk: A modified method of ligation-mediated genome walking for plant species with large genomes. *Analytical Biochemistry* **388**, 158-160.

U.S. Fish and Wildlife Service ESP (1996) Distinct Population Segment Policy.

Utami-Atmoko S, Singleton I, Van Noordwijk M, Van Schaik C, Mitra Setia T (2009) Male-male relationships in orangutans. In: *Orangutans: geographic variation in behavioral ecology and conservation.* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 225-234. Oxford University Press, New York.

Utami SS, Goossens B, Bruford MW, de Ruiter JR, van Hooff JA (2002) Male bimaturism and reproductive success in Sumatran orang-utans. *Behavioral Ecology* **13**, 643-652.

Vallender EJ, Mekel-Bobrov N, Lahn BT (2008) Genetic basis of human brain evolution. *Trends in Neurosciences* **31**, 637-644.

van Bers NEM, Van Oers K, Kerstens HHD*, et al.* (2010) Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. *Molecular Ecology* **19**, 89-99.

Van der Auwera GA, Carneiro MO, Hartl C*, et al.* (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

van Heel DA, Franke L, Hunt KA*, et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* **39**, 827-829.

van Noordwijk M, Sauren S, Nuzuar AA*, et al.* (2009) Development of independence: Sumatran and Bornean orangutans compared. In: *Orangutans: geographic variation in behavioral ecology and conservation.* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 189-203. Oxford University Press, Oxford.

van Noordwijk MA, Arora N, Willems EP*, et al.* (2012) Female philopatry and its social benefits among Bornean orangutans. *Behavioral Ecology and Sociobiology* **66**, 823-834.

van Schaik C, Ancrenaz M, Djojoasmoro R*, et al.* (2009a) Orangutan cultures revisited. In: *Orangutans: geographic variation in behavioral ecology and conservation. Oxford University Press, New York* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 299-310.

van Schaik C, Priatna A, Priatna D (1995) Population estimates and habitat preferences of orangutans based on line transects of nests. In: *The neglected ape*, pp. 129-147. Springer.

van Schaik CP (1999) The socioecology of fission-fusion sociality in orangutans. *Primates* **40**, 69-86.

van Schaik CP (2004) *Among orangutans : red apes and the rise of human culture* Belknap of Harvard University Press, Cambridge, MA.

van Schaik CP (2013) The costs and benefits of flexibility as an expression of behavioural plasticity: a primate perspective. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **368**.

van Schaik CP, Ancrenaz M, Borgen G*, et al.* (2003) Orangutan Cultures and the Evolution of Material Culture. *Science* **299**, 102-105.

van Schaik CP, Damerius L, Isler K (2013) Wild Orangutan Males Plan and Communicate Their Travel Direction One Day in Advance. *PLoS ONE* **8**, e74896.

van Schaik CP, Marshall AJ, Wich SA (2009b) Geographic variation in orangutan behavior and biology. In: *Orangutans - Geographic Variation in Behavioral Ecology and Conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP). Oxford University Press

van Tassell CP, Smith TP, Matukumalli LK*, et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods* **5**, 247-252.

van Woerden JT, Willems EP, van Schaik CP, Isler K (2012) Large brains buffer energetic effects of seasonal habitats in catarrhine primates. *Evolution* **66**, 191-199.

Verschoor EJ, Langenhuijzen S, Bontjer I*, et al.* (2004) The phylogeography of orangutan foamy viruses supports the theory of ancient repopulation of Sumatra. *Journal of virology* **78**, 12712-12716.

Verstappen HT (1997) The effect of climatic change on southeast Asian geomorphology. *Journal of Quaternary Science* **12**, 413-418.

Vignaud P, Duringer P, Mackaye HT*, et al.* (2002) Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152-155.

Vila C, Walker C, Sundqvist AK*, et al.* Combined use of maternal, paternal and bi-parental genetic markers for the identification of wolf-dog hybrids. *Heredity* **90**, 17-24.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4**, e72.

von Koenigswald GH (1982) Distribution and evolution of the orang-utan, Pongo pygmaeus (Hoppius). In: *The orang-utan: its biology and conservation*, pp. 1-15. Dr W. Junk Publishers, The Hague.

Voris HK (2000) Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *Journal of Biogeography* **27**, 1153-1167.

Wall JD, Andolfatto P, Przeworski M (2002) Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**, 203-216.

Wallner B, Piumi F, Brem G, Muller M, Achmann R (2004) Isolation of Y chromosome-specific microsatellites in the horse and cross-species amplification in the genus Equus. *Journal of Heredity* **95**, 158-164.

Wallner B, Vogl C, Shukla P*, et al.* (2013) Identification of genetic variation on the horse y chromosome and the tracing of male founder lineages in modern breeds. *PloS one* **8**, e60015.

Walsh T, Shahin H, Elkan-Miller T*, et al.* (2010) Whole Exome Sequencing and Homozygosity Mapping Identify Mutation in the Cell Polarity Protein GPSM2 as the Cause of

Nonsyndromic Hearing Loss DFNB82. *The American Journal of Human Genetics* **87**, 90-94.

Wandeler P, Camenisch G (in prep.) Beyond YCATS - identifying paternal lineages by long template PCR.

Wang X, Sun X, Wang P, Stattegger K (2009) Vegetation on the Sunda Shelf, South China Sea, during the last glacial maximum. *Palaeogeography, palaeoclimatology, palaeoecology* **278**, 88-97.

Warren KS, Verschoor EJ, Langenhuijzen S*, et al.* (2001) Speciation and Intrasubspecific Variation of Bornean Orangutans, Pongo pygmaeus pygmaeus. *Molecular Biology and Evolution* **18**, 472-480.

Waters PD, Wallis MC, Graves JAM (2007) Mammalian sex - Origin and evolution of the Y chromosome and SRY. *Seminars in Cell & Developmental Biology* **18**, 389-400.

Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**, 256-276.

Wei W, Ayub Q, Chen Y*, et al.* (2013a) A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome research* **23**, 388-395.

Wei W, Ayub Q, Xue Y, Tyler-Smith C (2013b) A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic Science International: Genetics* **7**, 568-572.

Wei Y, Lin-Lee Y-C, Yang X*, et al.* (2006) Molecular cloning of Chinese hamster 1q31 chromosomal fragile site DNA that is important to mdr1 gene amplification reveals a novel gene whose expression is associated with spermatocyte and adipocyte differentiation. *Gene* **372**, 44-52.

Weigel D, Mott R (2009) The 1001 genomes project for Arabidopsis thaliana. *Genome Biol* **10**, 107.

Weingrill T, Willems EP, Zimmermann N, Steinmetz H, Heistermann M (2011) Species-specific patterns in fecal glucocorticoid and androgen levels in zoo-living orangutans (Pongo spp.). *General and Comparative Endocrinology* **172**, 446-457.

Whitlock JR, Heynen AJ, Shuler MG, Bear MF (2006) Learning Induces Long-Term Potentiation in the Hippocampus. *Science* **313**, 1093-1097.

Whittaker DJ, Morales JC, Melnick DJ (2007) Resolution of the Hylobates phylogeny: Congruence of mitochondrial D-loop sequences with molecular, behavioral, and morphological data sets. *Molecular phylogenetics and evolution* **45**, 620-628.

Whitten T (2000) *The ecology of Sumatra* Tuttle Publishing.

Wich S, De Vries H, Ancrenaz M*, et al.* (2009a) Orangutan life history variation. In: *Orangutans - Geographic Variation in Behavioral Ecology and Conservation* (eds. Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 65-75. Oxford University Press

Wich S, Usher G, Peters H*, et al.* (2014) Preliminary data on the highland Sumatran orangutans (Pongo abelii) of Batang Toru. In: *High Altitude Primates*, pp. 265-283. Springer.

Wich S, Utami-Atmoko S, Mitra Setia T, Djoyosudharmo S, Geurts M (2006) Dietary and Energetic Responses of Pongo abelii to Fruit Availability Fluctuations. *International Journal of Primatology* **27**, 1535-1550.

Wich S, Utami Atmoko S, Mitra Setia T, van Schaik C (2009b) *Orangutans - Geographic Variation in Behavioral Ecology and Conservation* Oxford University Press New York.

Wich SA, Gaveau D, Abram N*, et al.* (2012) Understanding the Impacts of Land-Use Policies on a Threatened Species: Is There a Future for the Bornean Orang-utan? *PLoS ONE* **7**, e49142.

Wich SA, Meijaard E, Marshall AJ*, et al.* (2008) Distribution and conservation status of the orang-utan (Pongo spp.) on Borneo and Sumatra: how many remain? *Oryx* **42**, 329-339.

Wich SA, Riswan J, Refisch J, Nellemann C (2011a) *Orangutans and the economics of sustainable forest management in Sumatra* United Nations Environment Programme, Norway: Birkeland Trykkeri AS.

Wich SA, Vogel ER, Larsen MD*, et al.* (2011b) Forest Fruit Production Is Higher on Sumatra Than on Borneo. *PLoS ONE* **6**, e21278.

Wickham H (2009) *ggplot2: elegant graphics for data analysis* Springer Science & Business Media.

Wiedmann R, Smith T, Nonneman D (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics* **9**, 81.

Williams M (2012) The ~73 ka Toba super-eruption and its impact: History of a debate. *Quaternary International* **258**, 19-29.

Williams MA, Ambrose SH, van der Kaars S*, et al.* (2009) Environmental impact of the 73ka Toba super-eruption in South Asia. *Palaeogeography, palaeoclimatology, palaeoecology* **284**, 295-314.

Williams MAJ, Ambrose SH, der Kaars Sv*, et al.* (2010) Reply to the comment on "Environmental impact of the ~73kya Toba super-eruption in South Asia" by M. A. J. Williams, S. H. Ambrose, S. van der Kaars, C. Ruehlemann, U. Chattopadhyaya, J. Pal, P. R. Chauhan [Palaeogeography, Palaeoclimatology, Palaeoecology 284 (2009) 295–314]. *Palaeogeography, Palaeoclimatology, Palaeoecology* **296**, 204-211.

Wilton JC, Matthews GM (1996) Polarised membrane traffic in hepatocytes. *Bioessays* **18**, 229-236.

Wright JD (2000) Global climate change in marine stable isotope records. In: *Quaternary Geochronology. Methods and Applications*, p. 433. American Geophysical Union Washington, USA.

Xu X, Arnason U (1996) The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. *Journal of Molecular Evolution* **43**, 431-437.

Xue Y, Prado-Martinez J, Sudmant PH*, et al.* (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242-245.

Xue Y, Wang Q, Long Q*, et al.* (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* **19**, 1453-1457.

Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular biology and evolution* **23**, 212-226.

Yannic G, Basset P, Büchi L, Hausser J, Broquet T (2012) Scale-specific sex-biased dispersal in the valais shrew unveiled by genetic variation on the Y chromosome, autosomes, and mitochondrial DNA. *Evolution* **66**, 1737-1750.

Yannic G, Basset P, Hausser J (2008) Phylogeography and recolonization of the Swiss Alps by the Valais shrew (Sorex antinorii), inferred with autosomal and sex-specific markers. *Molecular Ecology* **17**, 4118-4133.

Young AL, Abaan HO, Zerbino D*, et al.* (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome Research* **20**, 249-256.

Zauner C, Schneeweiss B, Kranz A*, et al.* (2000) Resting energy expenditure in short-term starvation is increased as a result of an increase in serum norepinephrine. *The American Journal of Clinical Nutrition* **71**, 1511-1515.

Zeng K, Charlesworth B (2011) The Joint Effects of Background Selection and Genetic Recombination on Local Gene Genealogies. *Genetics* **189**, 251-266.

Zhang YW, Ryder OA, Zhang YP (2001) Genetic divergence of orangutan subspecies (Pongo pygmaeus). *Journal of Molecular Evolution* **52**, 516-526.

Zhernakova A, Alizadeh Behrooz Z, Bevova M*, et al.* (2007) Novel Association in Chromosome 4q27 Region with Rheumatoid Arthritis and Confirmation of Type 1 Diabetes Point to a General Risk Locus for Autoimmune Diseases. *American Journal of Human Genetics* **81**, 1284-1288.

Zhi L, Karesh WB, Janczewski DN*, et al.* (1996) Genomic differentiation among natural populations of orang-utan (Pongo pygmaeus). *Current Biology* **6**, 1326-1336.

Reprints of some relevant co-authored publications

MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 310–327 doi: 10.1111/mec.13027

Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation

ALEXANDER NATER,* MAJA P. GREMINGER,* NATASHA ARORA,* CAREL P. VAN SCHAIK,* BENOIT GOOSSENS,†‡§ IAN SINGLETON,¶** ERNST J. VERSCHOOR,†† KRISTIN S. WARREN‡‡ and MICHAEL KRÜTZEN*

**Anthropological Institute & Museum, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland, †Organisms and Environment Division, School of Biosciences, Cardiff University, Museum Avenue, Cardiff CF10 3AX, UK, ‡c/o Sabah Wildlife Department, Danau Girang Field Centre, Kota Kinabalu, Sabah, Malaysia, §Sabah Wildlife Department, Wisma Muis, 88100 Kota Kinabalu, Sabah, Malaysia, ¶Foundation for a Sustainable Ecosystem (YEL), Jl. Wahid Hasyim No. 51/74, Medan 20154, Indonesia, **Sumatran Orangutan Conservation Programme (PanEco-YEL), Jl. Wahid Hasyim No. 51/74, Medan 20154, Indonesia, ††Department of Virology, Biomedical Primate Research Centre, PO Box 3306, 2280 GH, Rijswijk, The Netherlands, ‡‡College of Veterinary Medicine, School of Veterinary and Life Sciences, Murdoch University, 90 South Street, Murdoch, WA 6150, Australia*

Abstract

Investigating how different evolutionary forces have shaped patterns of DNA variation within and among species requires detailed knowledge of their demographic history. Orang-utans, whose distribution is currently restricted to the South-East Asian islands of Borneo (*Pongo pygmaeus*) and Sumatra (*Pongo abelii*), have likely experienced a complex demographic history, influenced by recurrent changes in climate and sea levels, volcanic activities and anthropogenic pressures. Using the most extensive sample set of wild orang-utans to date, we employed an Approximate Bayesian Computation (ABC) approach to test the fit of 12 different demographic scenarios to the observed patterns of variation in autosomal, X-chromosomal, mitochondrial and Y-chromosomal markers. In the best-fitting model, Sumatran orang-utans exhibit a deep split of populations north and south of Lake Toba, probably caused by multiple eruptions of the Toba volcano. In addition, we found signals for a strong decline in all Sumatran populations ~24 ka, probably associated with hunting by human colonizers. In contrast, Bornean orang-utans experienced a severe bottleneck ~135 ka, followed by a population expansion and substructuring starting ~82 ka, which we link to an expansion from a glacial refugium. We showed that orang-utans went through drastic changes in population size and connectedness, caused by recurrent contraction and expansion of rainforest habitat during Pleistocene glaciations and probably hunting by early humans. Our findings emphasize the fact that important aspects of the evolutionary past of species with complex demographic histories might remain obscured when applying overly simplified models.

Keywords: Approximate Bayesian Computation, demographic history, *Pongo* spp., population structure

Received 24 April 2014; revision received 24 November 2014; accepted 27 November 2014

Introduction

Patterns of DNA variation are the result of both adaptive and nonadaptive processes, and the debate about the relative importance of natural selection and random genetic drift in shaping genetic diversity within and among species is still ongoing (e.g. Hahn 2008; Nei *et al.* 2010). A common approach to detect signals of selection aims at identifying genomic regions that show marked deviations in DNA variation from a neutral equilibrium model (reviewed in Nielsen 2005). However, under certain demographic scenarios, such as population size

Correspondence: Alexander Nater, Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala Sweden.
Fax: +46 18 4716310; E-mail: alexander.nater@ebc.uu.se

changes or population subdivision, random genetic drift can result in similar deviations as selection (e.g. Teshima *et al.* 2006; Excoffier *et al.* 2009). Therefore, confounding effects of demographic processes can only be unravelled from selective signals if the demographic history is explicitly taken into account when formulating the expectations under the neutral model against which observed patterns of DNA variation are tested (e.g. Haddrill *et al.* 2005; Stajich & Hahn 2005). Consequently, methods to reconstruct the demographic history of natural populations have recently generated great interest among evolutionary geneticists, as recent technical advances allow conducting genomewide studies of selection in a large variety of species (reviewed in Ellegren 2014).

Orang-utans, currently restricted to two distinct species on Borneo (*Pongo pygmaeus*) and northern Sumatra (*Pongo abelii*) (Wich *et al.* 2008), are the only Asian great apes and are phylogenetically most distant to humans (Groves 2001). Their ancestral position in the lineage leading to African great apes and modern humans has evoked great interest in this taxon in the overall effort to reconstruct the adaptive evolutionary history of great apes in general and humans in particular (Locke *et al.* 2011; Prado-Martinez *et al.* 2013). However, orang-utans might have experienced a complex demographic history, as their distribution has been subject to major changes during the Pleistocene. The ancestors of extant orang-utans have sequentially colonized the islands of the Sunda archipelago arriving from the South-East Asian mainland (Rijksen & Meijaard 1999; Delgado & Van Schaik 2000). Since then, their population history was strongly influenced by geological and climatic events: rising and falling sea levels cyclically connected and isolated the islands of Sundaland, allowing for potential terrestrial migration between the islands at certain points in time (Voris 2000).

Major volcanic eruptions, mainly on Sumatra and Java, might have led to the extinction of local orang-utan populations and subsequent recolonizations (Muir *et al.* 2000). Of special interest here is the Toba volcano on northern Sumatra, which has seen at least four major eruptions during the last 1.2 million years (Chesner *et al.* 1991). This sequence of eruptions culminated in the Toba supereruption ~73 ka, which is considered to be the most powerful volcanic eruption within the last 25 million years (Chesner *et al.* 1991) and is thought to have had severe consequences for flora and fauna on Sundaland (Williams *et al.* 2009). In the Late Pleistocene, all orang-utan populations on the mainland, southern Sumatra and Java went extinct (Rijksen & Meijaard 1999; Delgado & Van Schaik 2000). Climatic changes during the Pleistocene might have been responsible for the southward shift of the distribution and the disappearance of orang-utans from the mainland (Jablonski 1998). Moreover, anthropogenic factors, such as prehistoric hunting by hunter-gatherer societies, are likely to have played a significant role in the decline and extinction of orang-utans populations on insular South-East Asia (Delgado & Van Schaik 2000).

Genetic signals of these past demographic changes have been found in studies of genetic diversity in extant orang-utan populations on Borneo and Sumatra. Most genetic studies analysing autosomal and mitochondrial DNA (mtDNA) agree that Sumatran orang-utans show a higher level of sequence diversity and corresponding long-term effective population size (N_e) (Muir *et al.* 2000; Zhang *et al.* 2001; Steiper 2006; Locke *et al.* 2011; Prado-Martinez *et al.* 2013), even though Sumatran orang-utans have a much smaller current census size and a more restricted distribution than Borneans (~6600 vs. ~54 000 individuals, Wich *et al.* 2008). The large N_e of the Sumatran species was interpreted as a signal of immigration from multiple differentiated populations into the current Sumatran gene pool (Muir *et al.* 2000; Steiper 2006). However, Y-chromosomal diversity in orang-utans shows the opposite pattern compared to mtDNA and autosomal data, with a smaller N_e on Sumatra than Borneo (Nater *et al.* 2011). Such contrasting patterns of N_e between species and among genomic regions hint at complex population dynamics that have so far not been properly investigated.

Recently, Locke *et al.* (2011) used extensive single nucleotide polymorphism (SNP) data from whole-genome resequencing of five Bornean and five Sumatran orang-utans to model the demographic history of the two species. They found that a model with a population split ~400 ka with subsequent gene flow between Borneo and Sumatra fits the observed data best. Furthermore, Locke and colleagues inferred that Sumatran orang-utans underwent a continuous exponential population growth since the population split, while Bornean orang-utans were subject to a continuous exponential decline. Given the large amount of genetic data, the study by Locke and colleagues is currently regarded as the most accurate reconstruction of demographic history in orang-utans to date. However, the demographic modelling approach by Locke and colleagues did not take several idiosyncrasies of orang-utan biology into account, thus severely limiting the conclusions that could be drawn from their findings.

First, it has been shown that biased sampling and disregard of population structure will produce misleading results regarding N_e and its temporal changes (Stadler *et al.* 2009; Chikhi *et al.* 2010). The study by Locke and colleagues incorporated data from only five captive individuals from Borneo and Sumatra each without further provenance information. This limited genetic

sampling is unlikely to represent the entire genetic diversity present on both islands. Second, given the lack of detailed sample provenance, analyses were restricted to models treating Bornean and Sumatran orang-utans as single panmictic populations each. Previous work, however, unequivocally showed that both Bornean and Sumatran orang-utans are genetically deeply structured (Warren *et al.* 2001; Arora *et al.* 2010; Nater *et al.* 2011). In particular, on Sumatra, populations north and south of Lake Toba exhibit high genetic differentiation (Nater *et al.* 2011, 2013). Third, Locke and colleagues did not test complex demographic models including population bottlenecks or recent declines, as suggested in previous genetic studies. For example, genetic signals of a bottleneck with subsequent population expansion on Borneo might be linked to a glacial refugium or the impact of the Toba supereruption ~73 ka (Steiper 2006; Arora *et al.* 2010), and patterns of a recent population decline in Sabah, Borneo, are most likely attributable to recent anthropogenic pressures (Goossens *et al.* 2006).

Reconstructing the demographic history of a species has long been hindered by the fact that full-likelihood methods were restricted to relatively simple demographic models (e.g. Wilson *et al.* 2003; Hey & Nielsen 2004), which might not capture all relevant processes in complex demographic settings. This restriction is mainly caused by the fact that the computation of the likelihood function of complex demographic models with many parameters is either intractable or computationally too expensive, especially for large data sets (Marjoram *et al.* 2003). Approximate Bayesian Computation (ABC) allows circumventing these problems by approximating the likelihood functions with simulations of genetic data under a given demographic model (Beaumont *et al.* 2002; Marjoram *et al.* 2003). To estimate the model parameters, parameter values are drawn from predefined prior distributions and used to simulate genetic data matching the observed data in type of markers and number of loci. Both observed and simulated data are then reduced to a set of summary statistics, and the Euclidian distance between the observed and the simulated summary statistics is calculated. Based on the subset of simulations with the smallest Euclidian distance between observed and simulated data, the posterior distribution of the model parameters can be approximated and the relative fit of different demographic models to the data can be assessed.

Here, we present an ABC modelling approach of the demographic history of orang-utans based on autosomal and sex-linked marker systems. We aim to improve the current knowledge of demographic history by applying three major improvements over previous studies. First, we capitalize on the knowledge base of behavioural ecology and population genetics of orang-utans to test realistic demographic models. Second, due to our extensive set of orang-utan samples with detailed and reliable provenance, we are able to investigate models incorporating population substructure in both orang-utan species, which allows us to disentangle changes in population size from confounding effects due to changes in population structure. Third, by combining autosomal and sex-linked markers into a single demographic analysis, we take advantage of the specific information content of different marker systems in this species with its heavily sex-biased dispersal. Due to strong female philopatry in orang-utans (Galdikas 1995; Arora *et al.* 2012; van Noordwijk *et al.* 2012), mitochondrial markers contain information about population split times without the confounding influence of gene flow. In contrast, Y-chromosomal loci should have more power than autosomal markers to reveal low levels of male-mediated gene flow.

Materials and methods

Sample collection and genetic markers

A representative sampling scheme covering the whole range of a species is crucial for accurate reconstruction of demographic history (Stadler *et al.* 2009). We used an extensive set of samples from wild-born orang-utans from 10 sampling locations, covering the entire distribution of the genus (Table 1 and Fig. 1, see Supporting information for detailed information about sample origin). Samples were analysed for several genetic marker systems with different modes of inheritance and effective population sizes (Table 2), thus ensuring representation of both male and female population history, an important aspect in demographic reconstructions in species with strongly sex-biased dispersal (Nater *et al.* 2011; Nietlisbach *et al.* 2012).

The autosomal microsatellite data contained genotypes of 25 microsatellite markers from a total of 237 individuals (Arora *et al.* 2010; Nater *et al.* 2013; Greminger *et al.* 2014). We also included sequences from three mtDNA genes with a total length of 1355 bp from 118 individuals (Nater *et al.* 2011), and Y-chromosomal haplotypes based on 11 Y-linked microsatellite loci from 129 individuals (Nater *et al.* 2011; Nietlisbach *et al.* 2012). We complemented the data set by additionally sequencing 8055 bp of the noncoding X-chromosomal region Xq13.3 (Kaessmann *et al.* 2001) in 36 individuals and four noncoding autosomal regions (Fischer *et al.* 2006) of a total of 8238 bp in 22 individuals. Basic summary statistics for all marker systems are provided in Table 2. The primers and cycling conditions used for PCR amplification and sequencing of the autosomal and X-chromosomal regions are

Table 1 Sample sizes for the different marker systems in the 10 geographic regions

Sampling region*	mtDNA	Y-STRs	Autosomal STRs	Autosomal regions	Xq13.3
North Kinabatangan (NK)	6	10	32	4	3
South Kinabatangan (SK)	13	15	76	4	3
East Kalimantan (EK)	7	9	34	4	5
Sarawak (SR)	8	2	12	2	1
Central Kalimantan (CK)	9	9	68	2	2
West Kalimantan (WK)	9	8	32	4	4
Batang Toru (BT)	8	8	18	4	3
North Aceh (NA)	7	15	32	6	3
Langkat (LK)	14	15	66	10	6
West Alas (WA)	37	38	104	4	7
Total	118	129	474	44	37

Sample sizes are given as number of sampled chromosomes. The light grey shading refers to Bornean populations, middle grey to Sumatran populations north of Lake Toba and dark grey to the Sumatran population south of Lake Toba.
*Sampling regions corresponding to Fig. 1.

described in the Supporting Table S1, Supporting information.

Approximate Bayesian Computation

Model selection procedure. We reconstructed the demographic history of orang-utans using an ABC approach implemented in the software package ABCTOOLBOX v1.1 (Wegmann *et al.* 2010). To achieve this goal, we first performed a model selection procedure, in which we used a hierarchical approach to test a total of 12 different demographic models (Fig. 2) with increasing levels of complexity (see Tables S5 and S6, Supporting information, for more details about model parameterization and prior distributions).

We started by testing four relatively simple models assuming a single population for each of the two orang-utan species (Fig. 2A). The first model in this set (I2) posited constant population sizes and no migration between the two populations. The second model (IM2) incorporated asymmetric migration after the population split, up to a point in the past where migration between Borneo and Sumatra ceased. Gene flow in all models with migration was strictly male-mediated, as recent genetic and behavioural findings showed extreme female philopatric tendencies in orang-utans (Nater *et al.* 2011; Arora *et al.* 2012; van Noordwijk *et al.* 2012). The third model (IM2-GR) additionally allowed the two populations to change size exponentially after the population split and corresponded largely to the favoured model in the genomic study by Locke *et al.* (2011). In the fourth and most complex 2-population model (IM2-BN-GR), both populations retained a constant size after the population split, with the possibility for a sudden population size rescale followed by exponential growth or decline.

To test more biologically relevant demographic scenarios, we designed a series of 10-population models incorporating the repeatedly reported extensive population substructure in extant orang-utan populations (Warren *et al.* 2001; Goossens *et al.* 2005; Kanthaswamy *et al.* 2006; Arora *et al.* 2010; Nater *et al.* 2011, 2013). The use of 10 extant population units models is justified by previously published data (Arora *et al.* 2010; Nater *et al.* 2011, 2013; Greminger *et al.* 2014). The combination of patterns of population differentiation in both mtDNA and autosomal microsatellite markers points to six populations on Borneo, one Sumatran population south of Lake Toba and three Sumatran populations north of Lake Toba (see validation of population units in Supporting information). For all 10-population models, we assumed equal population sizes and equal symmetric migration rates among all populations within Borneo and among all populations north of Lake Toba, respectively, as well as a separate population size parameter for the population south of Lake Toba. We included asymmetric migration rates between Borneo and south of Lake Toba, and between north of Lake Toba and south of Lake Toba.

To assess to what extent the additional population units improve model fit, we first tested the best-fitting 2-population model against two basic 10-population models (IM10 and $IM10_{BO\text{-}NT}$, Fig. 2B). The IM10 model incorporated the population splitting sequence derived from mtDNA data, that is the populations north and south of Lake Toba show the oldest split, while Bornean populations diverged after this split (Nater *et al.* 2011). As this is in discordance with the current species designation (Groves 2001), which assigns a single species each to Sumatra and Borneo, we also tested this model against a model following the species split pattern ($IM10_{BO\text{-}NT}$), that is with the oldest split between

Table 2 Summary statistics for the marker systems used in the ABC analysis

Sequences	L_{Bases}*	Group	N_{Ind}†	N_{Seg}‡	π§	θ_W¶	D**
mtDNA (16S, ND3, CYTB)	1355	Borneo	52	19	0.0022	0.0031	−0.92
		South Toba	8	1	0.0002	0.0003	−1.05
		North Toba	58	41	0.0100	0.0066	1.79
Autosomal regions (Chr2a_R17, Chr9_R16, Chr12_R1, Chr19_R7)	8238	Borneo	10	19.50 ± 4.56	0.0033 ± 0.0011	0.0027 ± 0.0006	0.68 ± 0.66
		South Toba	2	13.50 ± 7.79	0.0037 ± 0.0020	0.0036 ± 0.0020	0.08 ± 0.49
		North Toba	10	28.75 ± 4.66	0.0046 ± 0.0012	0.0040 ± 0.0006	0.51 ± 0.52
Xq13.3	8055	Borneo	18	6	0.0001	0.0002	−1.11
		South Toba	3	33	0.0027	0.0027	0.00
		North Toba	15	54	0.0020	0.0020	−0.09

Microsatellites	N_{Loci}††	Group	N_{Ind}	N_A‡‡	H_O§§	H_E¶¶	G-W***
Autosomal STR	25	Borneo	127	7.16 ± 4.13	0.53 ± 0.22	0.61 ± 0.25	0.90 ± 0.15
		South Toba	9	3.84 ± 1.18	0.60 ± 0.23	0.62 ± 0.16	0.72 ± 0.22
		North Toba	101	6.32 ± 3.11	0.61 ± 0.16	0.65 ± 0.16	0.82 ± 0.17
Y-STR	11	Borneo	53	3.18 ± 2.48	–	0.31 ± 0.33	0.90 ± 0.14
		South Toba	8	1.27 ± 0.65	–	0.08 ± 0.19	0.88 ± 0.18
		North Toba	68	1.91 ± 1.64	–	0.12 ± 0.24	0.91 ± 0.17

Statistics are provided as average and standard deviation for marker systems with multiple independent loci.
*Sequence length in base pairs.
†Number of sampled individuals.
‡Number of segregating sites
§Nucleotide diversity
¶Watterson's θ per base pair.
**Tajima's D (Tajima 1989).
††Number of loci.
‡‡Number of alleles.
§§Observed heterozygosity.
¶¶Expected heterozygosity.
***Garza–Williamson index (Garza & Williamson 2001).

Sumatra and Borneo, to see whether incomplete lineage sorting could be responsible for the particular phylogenetic pattern observed for mtDNA.

We further tested for the presence of population size changes in the demographic history of orang-utans, as suggested by previous studies (Goossens *et al.* 2006; Steiper 2006; Arora *et al.* 2010; Locke *et al.* 2011). First, we tested for signals of recent declines in Sumatra (IM10-DEC_{SU}), Borneo (IM10-DEC_{BO}) or both islands (IM10-DEC_{ALL}) (Fig. 2C).

In a second test, we evaluated the support for a bottleneck on Borneo (IM10-BN_{BO}-DEC_{SU}), possibly linked to a refugium during the penultimate glaciation (Arora *et al.* 2010) (Fig. 2D).

Last, we tested for evidence for a bottleneck on Sumatra linked to the Toba supereruption, either allowing for a broad prior range of the magnitude of decline (IM10-BN_{BO}-TOBA-DEC_{SU}) or restricting to a severe bottleneck of <100 individuals in each of the four Sumatran populations (IM10-BN_{BO}-RECOL-DEC_{SU}), resembling a founder effect after local extinction and recolonization events on Sumatra (Fig. 2E).

ABC data simulation. To simulate genetic data under different demographic models, we used the software FASTSIMCOAL v1.1.2 (Excoffier & Foll 2011). Simulations for the different marker systems were run with the same set of parameters, whereby the effective population sizes were scaled 1 to 0.75, 0.25 and 0.25 for autosomal, X-chromosomal, mitochondrial and Y-chromosomal markers, respectively. We then used ARLSUMSTAT v3.5.1.3 (Excoffier & Lischer 2010) to calculate a total of 259 summary statistics for each simulated data set as well as for the observed data set (Table S7, Supporting information). The summary statistics were chosen to capture the information in the genetic data about population differentiation, within population diversity, and population size changes. To avoid problems with unreliable phasing, we only used summary statistics that do not require phased sequence data for X-chromosomal and autosomal loci. As the number of simulated populations differed between the 2-population and 10-population models, summary statistic would not be directly comparable between the two sets of models. Therefore, when running the 10-population models, we applied a script

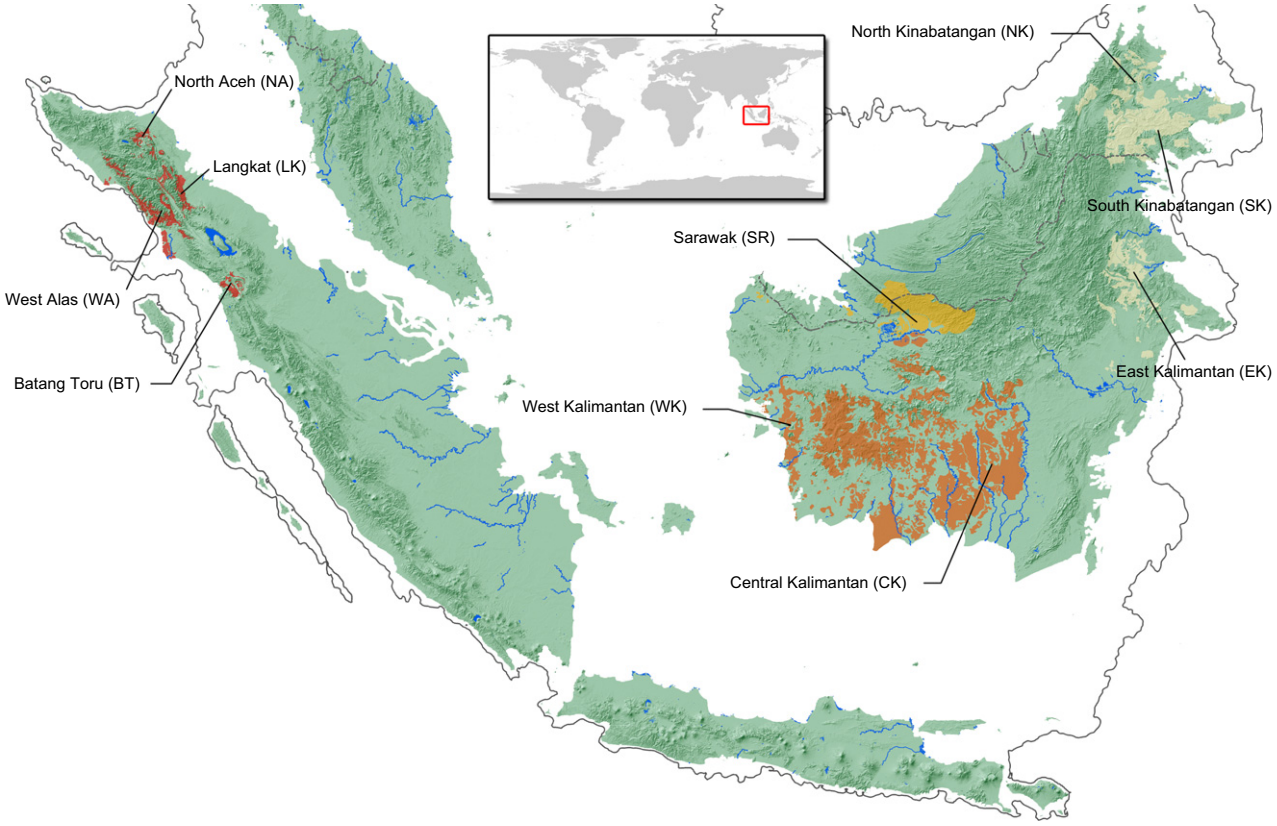


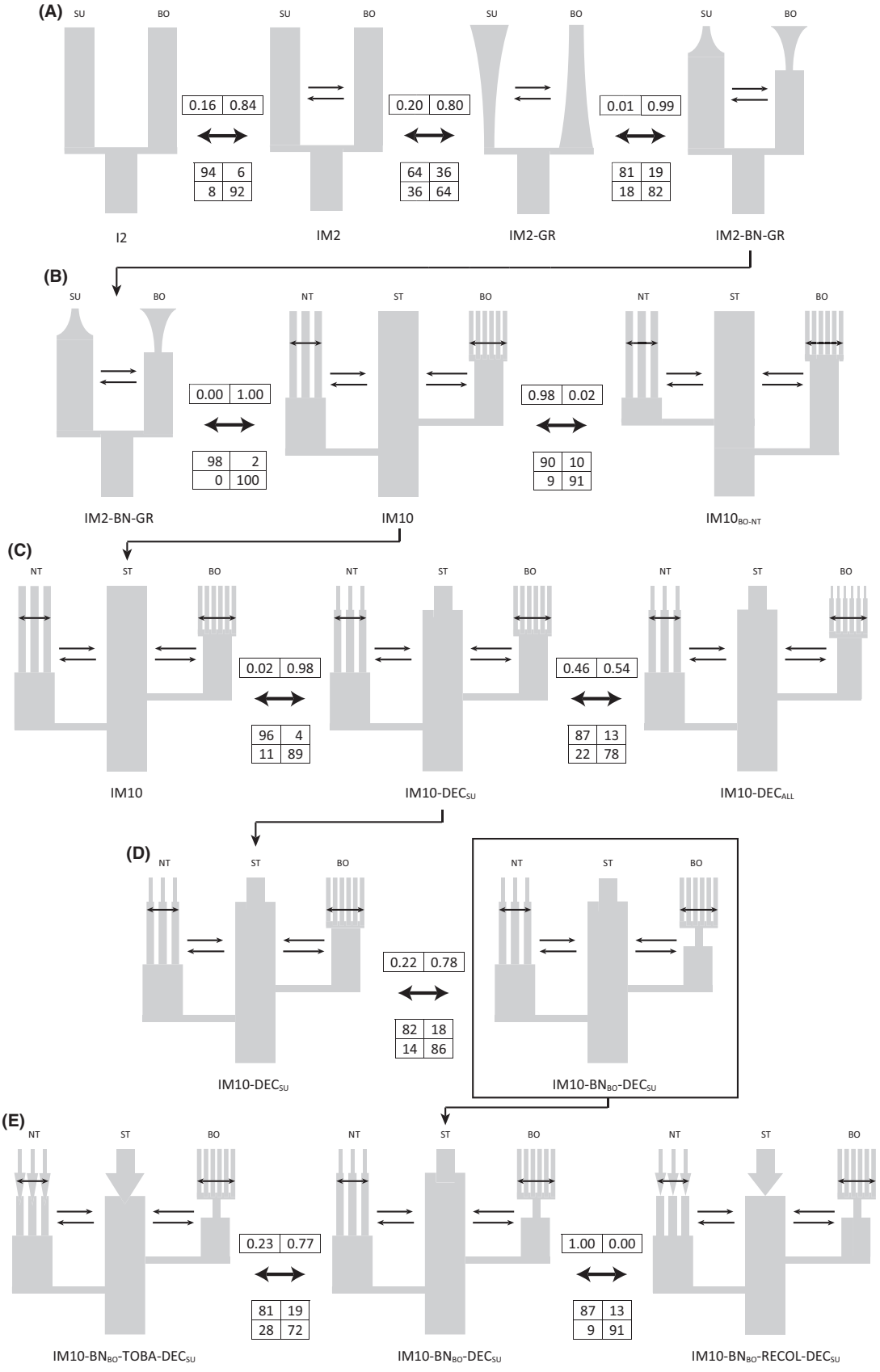
Fig. 1 Map of sampling regions in Sundaland used for the demographic modelling. Shaded areas represent the current distribution of the Sumatran orang-utans and the three subspecies of Bornean orang-utans. The grey line indicates the extent of the exposed Sunda shelf during the last glacial maximum (19–26 ka, 120 m below current sea level).

pooling the simulated data into a Bornean and a Sumatran group after each simulation step. Summary statistics were then also calculated islandwise, in order to be able to directly compare to the 2-population models.

We first performed an initial run of 2×10^6 simulations with the standard rejection sampler (Tavare *et al.* 1997). These simulations were used for both model selection and validation. To reduce the dimensionality of the summary statistics, we performed a principal component analysis (PCA) with the 'prcomp' function in R version 2.12.1 (R Development Core Team 2010). We pooled and standardized 100 000 random simulations from each of the two compared models and used these summary statistics to extract the loadings of the first 10 principal components. We then transformed both the simulated and the observed data to perform a multinomial logistic regression with the R package 'abc' version 1.6. For this, we used the 0.1% of the simulations with the smallest Euclidean distance between the transformed summary statistics and the observed data.

To assess model fit, we also calculated the marginal densities and the probability of the observed data under the general linear model (GLM) used for the postsampling regression for each model with ABCTOOLBOX (Leuenberger & Wegmann 2010). For this, we again transformed both the simulated summary statistics and the observed data with the loadings for the first 10 principal components. This time, PCA loadings were obtained for each model separately using 100 000 random simulations. The GLM was built from the 2000 simulations closest to the observed data, and we assessed the goodness of fit of all tested models to the observed data by calculating the *P*-value of the observed data under the GLM (Supporting Table S8, Supporting information). The *P*-value is representing the proportion of the retained simulations showing a lower or equal likelihood under the inferred GLM as compared to the observed genetic data (Wegmann *et al.* 2009b). Thus, low *P*-values indicate that the observed data is unlikely to have been generated under the inferred GLM, implying a bad model fit.

Parameter estimation. To obtain good estimates of the posterior distributions of the parameters for the best-fitting model (IM10-BN_{BO}-DEC_{SU}), we used a MCMC



without likelihood method (Wegmann *et al.* 2009a). To reduce the dimensionality of the data and extract as much information as possible about the model parameters, we used the first 20 000 simulations with the standard sampler to define the first 12 orthogonal components of the summary statistics that maximize the covariance matrix between summary statistics and model parameters. For this, we applied a partial least-squares (PLS) regression approach (Boulesteix & Strimmer 2007) as implemented in the 'pls' R package (Mevik & Wehrens 2007) and used the R script provided in the ABCTOOLBOX package. We defined the optimal number of PLS components by assessing the drop in the root-mean-squared error for each parameter with the inclusion of additional PLS components. This way, a large set of summary statistics is reduced to a number of independent components, whereby summary statistics that are most informative about the model parameters are weighted more than summary statistics that do not show much response to changing parameter values (Wegmann *et al.* 2009a). The initial simulations were also used to define the tolerance distance based on a tolerance level of 0.1 and to calibrate the transition kernel of the MCMC run with a rangeProp setting of 1 unit of standard deviation (Wegmann *et al.* 2009a, 2010). We then ran a total of 10^7 iterations with the MCMC sampler, followed by a ABC-GLM postsampling regression (Leuenberger & Wegmann 2010) on the 10 000 simulations with the smallest Euclidean distance to the PLS components of the observed summary statistics. Finally, we used R to plot the posterior distributions of important model parameters.

ABC validation. The performance of ABC in model selection and parameter estimation in complex demographic settings inevitably suffers from the loss of information when the observed and simulated genetic data are reduced to a set of summary statistics (Robert *et al.* 2011). This necessitates a careful validation of the employed ABC procedure to avoid biases in the approximation of posterior probabilities of evaluated models and the estimation of model parameters. Accordingly, we validated our model selection and parameter estimation approach with four different procedures. The first three validation approaches made use of so-called pseudo-observed data sets (*pods*), whereby parameter combinations are randomly drawn from the prior distributions and the corresponding summary statistics were simulated under a given model. These sets of summary statistics were then treated as if it were real observed data, but as the model and the corresponding parameter values that generated these summary statistics were known, we could use the *pods* to validate both our model selection and parameter estimation procedure.

In the first validation step, we investigated the model misclassification rate for each pairwise model comparison by generating 100 *pods* under each model with parameters randomly drawn from the prior distributions. We then performed the same model selection procedure as with the real observed data and counted the number of assignments to each model. We derived the model misclassification rate by counting all assignments of *pods* to a model other than the one generating it (Fig. 2).

Second, we assessed the accuracy of the parameter estimation, in terms of both different point estimators (mode, average and median) and over the whole posterior distribution under different tolerance levels (proportion of retained simulations). For this, we generated 1000 *pods* under the best-fitting model (IM10-BN_{BO}-DEC_{SU}) and performed the same parameter estimation procedure on each *pods* as for the real data. The accuracy of the point estimators was assessed using the average of the root-mean-squared errors (RMSE) over all 1000 *pods* (Table S9, Supporting information), while the root-mean-integrated-squared error (Leuenberger & Wegmann 2010) was used to assess accuracy over the whole posterior distribution (Table S10, Supporting information). The results indicated that accuracy of the posterior distributions

Fig. 2 Schematic representation of the hierarchical model testing procedure. The 12 tested demographic models can be divided into four 2-population models and eight 10-populations models (IM10-DEC_{BO} not shown). The box above the left–right arrow shows the model posterior probabilities for each model comparison pair. The overall best-fitting model (IM10-BN_{BO}-DEC_{SU}) is shown in a black frame. The box below the left–right arrow shows the power to distinguish between the two compared models as evaluated in a cross-validation procedure with 100 validations for each model, with the upper left and lower right boxes showing the correct model assignments for model 1 and model 2, respectively (SU = Sumatra, BO = Borneo, NT = Sumatra north of Lake Toba, ST = Sumatran south of Lake Toba). (A) Comparison of four 2-population models, testing gene flow after the population split, exponential population growth or decline after the population split and sudden population size change followed by exponential growth or decline. (B) Comparison between the best-fitting 2-population model and two 10-population models incorporating population structure. (C) Tests of recent population declines in Sumatra, and Sumatra as well as Borneo. (D) Test of population bottleneck on Borneo. (E) Testing of a population bottleneck on Sumatra associated with the Toba supereruption 65–75 ka. The leftmost model implements a bottleneck in all four populations on Sumatra, followed by exponential population recovery. The rightmost model is similar, but restricts the bottleneck to a size of <100 surviving individuals per population, thus representing a scenario where regions devastated by the Toba eruption were recolonized from other areas after restoration of the rain forest habitat.

is little affected by varying tolerance levels and that the mode of the distribution is the most accurate point estimator for parameter estimation.

Third, to increase confidence in the parameter estimates of the best-fitting model, we checked for biased posterior distributions by producing 1000 *pods* under the best-fitting model with parameter values drawn from the prior distributions. We used ABCTOOLBOX to calculate the posterior quantiles of the true parameter values within the estimated posterior distributions for each *pods* and used a Kolmogorov–Smirnov test for uniformity in R (Wegmann *et al.* 2009a). Significant deviation from uniformity after sequential Bonferroni correction (Rice 1989) would indicate biased posterior distributions (Cook *et al.* 2006). The distribution of posterior quantiles within which the true values of the *pods* fell did not significantly deviate from the expectation of uniformity for most parameters (Fig. S4, Supporting information). In most cases where the posterior quantiles were not distributed uniformly, data points were overrepresented in the centre of the histogram, indicating that the posterior distributions were estimated too conservatively.

In a last validation approach, we tested whether the best-fitting model (IM10-BN_{BO}-DEC_{SU}) and the corresponding posterior distributions of the model parameters are able to reproduce the summary statistics of the observed data. For this, we randomly sampled 10 000 parameter sets from the inferred posterior distributions and used these to simulate genetic data under the best-fitting model. We then carried out a PCA transformation of the simulated data and plotted the first 16 principal components to check whether the transformed observed data fell within the distribution of the simulated data (Fig. S5, Supporting information). This was the case for all the first 16 principal components, suggesting that the best-fitting model and its inferred parameter estimates are well able to explain the observed data.

Results

Model selection

We tested 12 demographic models, evaluating the impact of multiple demographic processes on the current genetic makeup of orang-utan populations (Fig. 2). We first compared simple models that treated Bornean and Sumatran orang-utans as single populations, but differed in the opportunity for migration after the population split (IM2 vs. I2, Fig. 2A). We found substantial support for the model allowing migration after the split [IM2, Bayes factor, i.e. ratio of model posterior probabilities (BF) 5.18]. However, this simple isolation with migration model achieved only a very poor fit to the observed data, as shown by the probability of the observed data under the GLM used for parameter estimation (GLM *P*-value) of 0.003, indicating that additional processes were involved in shaping the gene pool of orang-utans. Of all four 2-population models tested, we observed a very strong support for a model that allowed a sudden change in population size for both populations followed by exponential growth (IM2-BN-GR vs. I2, IM2, IM2-GR, BF 36.79). Still, this model did not achieve a good fit to the observed data, as evidenced by a *P*-value of the observed data under the GLM of only 0.017 (Table S8, Supporting information).

The poor model fit of all tested 2-population models can be explained by the extensive population substructure within the two orang-utan species (Warren *et al.* 2001; Kanthaswamy *et al.* 2006; Arora *et al.* 2010; Nater *et al.* 2011, 2013), which differs to a great extent for female- and male-mediated marker systems (Nater *et al.* 2011; Nietlisbach *et al.* 2012). Accordingly, the N_e for each marker system varies to a large degree and cannot be described accurately with just one population size parameter per island. In agreement with this notion, we found that a basic model with 10 current population units (IM10) achieved a better fit to the observed genetic data (GLM *P*-value 0.224) than all the 2-population models (Table S8, Supporting information), and also obtained much stronger statistical support when directly compared against the best 2-population model (IM10 vs. IM2-BN-GR, BF 830.21, Fig. 2B). However, in our case, a better fit of the 10-population model compared to the 2-population models was not unexpected, as part of the observed genetic data was used beforehand to derive the number of population units in the 10-population models. When we computed summary statistics for the IM10 model without pooling the genetic data for the Sumatran populations north and south of Lake Toba, the model fit was still poor (GLM *P*-value 0.019). To improve model fit, we first tested whether a population split sequence following the species designation fits the data better than the pattern suggested by mtDNA data (deepest split within Sumatran orang-utans north and south of Lake Toba). This was strongly rejected by the observed data (IM10 vs. $IM10_{BO\text{-}NT}$, BF 45.45, Fig. 2B).

We then further tested for recent population declines in Sumatra (IM10-DEC_{SU} vs. IM10, BF 57.03), on Borneo (IM10-DEC_{BO} vs. IM10, BF 0.48) or in both islands (IM10-DEC_{ALL} vs. IM10-DEC_{SU}, BF 0.94, Fig. 2C). Incorporating a population decline in Sumatra considerably improved the model fit (GLM *P*-value 0.553).

Next, we tested a model incorporating a bottleneck on Borneo together with a recent decline in Sumatra (Fig. 2D), which revealed substantial support for a bot-

tleneck on Borneo (IM10-BN_{BO}-DEC_{SU} vs. IM10-DEC_{SU}, BF 3.60).

Finally, we evaluated the statistical support for a bottleneck on Sumatra associated with the Toba supereruption (Fig. 2E). We found substantial support against a bottleneck on Sumatra in general (IM10-BN_{BO}-DEC_{SU} vs. IM10-BN_{BO}-TOBA-DEC_{SU}, BF 3.29) and overwhelming support against a severe bottleneck (less than 100 individuals per population) (IM10-BN_{BO}-DEC_{SU} vs. IM10-BN_{BO}-RECOL-DEC_{SU}, BF 10 887.60).

After performing a series of hierarchical model selection steps, we were able to identify a demographic model (IM10-BN_{BO}-DEC_{SU}) capable of reproducing the observed patterns of DNA variation in the two current orang-utan species. Therefore, this model is likely to capture the biologically most relevant processes in the demographic history of orang-utans.

Parameter estimation

We estimated the model parameters for the selected 10-population model (IM10-BN_{BO}-DEC_{SU}, Fig. 3) based on a total of 10 million simulations (Table 3, Fig. 4). The parameter estimates point to a current N_e of ~970 diploid individuals in each of the six Bornean populations. We found support for a bottleneck on Borneo starting ~135 ka and ending ~82 ka, during which N_e on Borneo was reduced from an ancestral N_e of ~17 000 individuals to ~2600 individuals. The bottleneck on Borneo was followed by population recovery and substructuring, with a current total N_e of all Bornean populations of ~6150 individuals.

On Sumatra, the three populations north of Lake Toba suffered a decline ~24 ka from a N_e of ~10 500 to currently only ~960 individuals in each of the three populations, corresponding to a total N_e in the metapopulation north of Lake Toba of ~38 300 and ~3300 individuals before and after the decline, respectively. We estimated that population structure north of Lake Toba was established ~860 ka, with an ancestral effective population size of ~14 400 individuals. The population south of Lake Toba also went through a recent decline ~24 ka from a N_e of ~24 200 individuals in the ancestral population to currently only ~1030 individuals. Thus, Sumatran orang-utan populations first expanded during the Middle Pleistocene before experiencing an islandwide population crash in the Late Pleistocene or Early Holocene.

We inferred the population split time between Borneo and south Toba as ~1.13 Ma, and between north and south of Lake Toba as ~3.39 Ma. Gene flow between Borneo and Sumatra appears to have ceased ~87 ka, but this parameter was associated with a broad posterior distribution. We found no evidence for asymmetric migration rates between Borneo and south of Lake Toba, and between south of Lake Toba and north of Lake Toba. The migration rates between the two islands were comparable to the migration rates over the Toba caldera on Sumatra, while migration rates among the populations on Borneo and among those north of Toba, respectively, were estimated to be about a magnitude higher.

Discussion

Our modelling approach capitalized on the use of multiple genetic marker systems and an extensive set of geographically well-defined samples, in contrast to

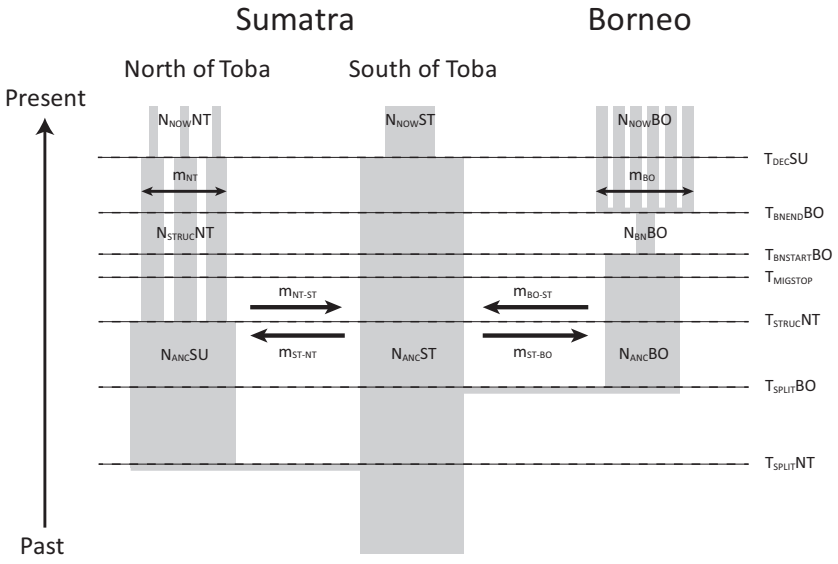


Fig. 3 Schematic representation of the selected 10-population model with a bottleneck on Borneo and recent population declines in all Sumatran populations (IM10-BN_{BO}-DEC_{SU}).

Table 3 Estimates of the model parameters for the selected 10-population model with a bottleneck on Borneo and a recent decline in Sumatra (IM10-BN_{BO}-DEC_{SU})

Parameter	Prior*	Mode	Mean	90%-HPD†
N_{NOW}BO [ind] (6)	logunif[100;10 000]	974	1028	348–3011
N_{NOW}NT [ind] (3)	logunif[100;10 000]	963	933	239–3613
N_{NOW}ST [ind] (1)	logunif[100;10 000]	1034	952	189–4514
N_{BN}BO [ind] (1)	logunif[100;10 000]	2598	1486	286–9988
N_{ANC}BO [ind] (1)	logunif[1000;100 000]	17 046	12 344	2171–89 115
N_{STRUC}NT [ind] (3)	logunif[1000;100 000]	10 508	11 278	1886–78 264
N_{ANC}NT [ind] (1)	logunif[1000;100 000]	14 407	10 519	1565–70 259
N_{ANC}ST [ind] (1)	logunif[1000;100 000]	24 193	13 991	2629–99 070
T_{BNEND}BO [years]	unif[8750;400 000]	81 946	149 580	8848–283 785
$T_{BNSTART}$BO [years]	T_{BNEND}BO + unif[250;100 000]	135 076	191 001	20 855–348 145
T_{SPLIT}BO [ka]	unif[400;1500]	1128	960	497–1436
T_{DEC}SU [years]	unif[1.0;3.5]	23 651	36 200	4119–67 272
T_{STRUC}NT [ka]	unif[75;1500]	861	820	267–1398
T_{SPLIT}NT [ka]	unif[1500;4000]	3392	2995	2101–3999
$T_{MIGSTOP}$ [years]	unif[2.5;4.2]	87 034	161 862	8849–310 833
Log($m_{BO\text{-}ST}$) [migrants/ind/gen]	unif[−5.0; −3.0]	−3.55	−3.96	−4.79 to −3.09
Log($m_{ST\text{-}BO}$) [migrants/ind/gen]	unif[−5.0; −3.0]	−3.42	−3.84	−4.61 to −3.10
Log($m_{NT\text{-}ST}$) [migrants/ind/gen]	unif[−5.0; −3.0]	−3.89	−3.98	−4.81 to −3.14
Log($m_{ST\text{-}NT}$) [migrants/ind/gen]	unif[−5.0; −3.0]	−3.65	−3.92	−4.71 to −3.06
Log(m_{BO}) [migrants/ind/gen]	unif[−4.0; −2.0]	−2.52	−2.90	−3.66 to −2.02
Log(m_{NT}) [migrants/ind/gen]	unif[−4.0; −2.0]	−2.51	−2.89	−3.65 to −2.03

BO, Borneo, NT, Sumatra north of Lake Toba; ST, Sumatra south of Lake Toba; N_{NOW}, current effective population size; N_{BN}, effective population size during population bottleneck; N_{ANC}, ancestral effective population size; N_{STRUC}, effective population size before recent decline; T_{BNEND}, time since population bottleneck ended; $T_{BNSTART}$, time when population bottleneck started; T_{SPLIT}, population split time; T_{DEC}, time since population decline; T_{STRUC}, time since establishment of population structure; $T_{MIGSTOP}$, time since migration between Borneo and Sumatra stopped; m, migration rate per individual per generation (an illustration of the meaning of the different model parameters can be found in Fig. 3), the number in parentheses next to the population size parameters refer to the number of simulated populations of this size each.

*The prior distributions for the parameter values were either uniform or loguniform within the boundaries provided in squared brackets

†Ninety percent highest posterior density interval.

previous studies, which based their findings on a small number of captive individuals with poorly recorded provenance (Locke *et al.* 2011; Mailund *et al.* 2011, 2012). Thus, our study was able to shed light on important aspects of orang-utan demographic history that so far remained unexamined due to nonrepresentative sampling and dismissal of within-species population structure. For instance, the inferred model by Locke *et al.* (2011) of a continuously expanding Sumatran orang-utan population with a substantially larger current N_e as compared to Bornean orang-utans was unrealistic in the light of current species distribution and abundance and did not capture recent population dynamics. Our results indicate that such misleading signals are the result of a recent decline and deep divergence of orang-utan populations on Sumatra, which yields a larger long-term N_e for Sumatran orang-utans as compared to Bornean orang-utans in oversimplified demographic models.

Inference of best-fitting model

We inferred that a model with comprehensive population structure, a bottleneck on Borneo and a recent decline in Sumatra (IM10-BN_{BO}-DEC_{SU}), fits the observed data significantly better than a range of simplified models that treat each orang-utan species as a single panmictic population. Estimation of demographic parameters under this model revealed a population split time between Borneo and Sumatran populations south of Lake Toba of just over a million years ago, followed by bidirectional gene flow. This species split time estimate is considerably older than estimates obtained using whole-genome data, suggesting a species split time of between 330 and 600 ka (Locke *et al.* 2011; Mailund *et al.* 2011, 2012). Such recent species split estimates are, however, in disagreement with findings based on mitochondrial DNA, which yielded divergence time estimates of island-specific mtDNA lineages of 1–5 Ma (Xu & Arnason 1996; Zhi *et al.* 1996;

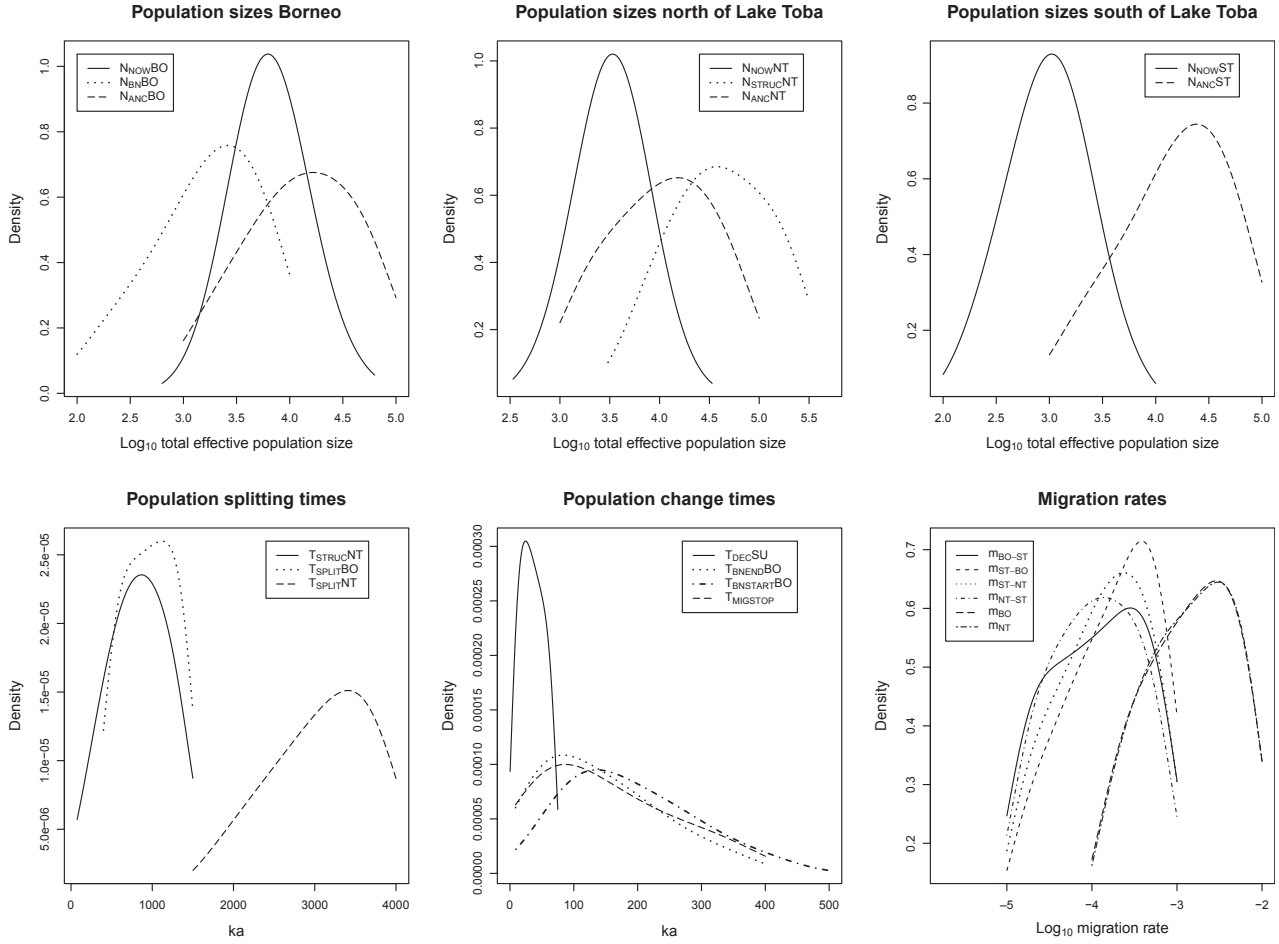


Fig. 4 Posterior distributions of important model parameter under the selected 10-population model (IM10-BN_{BO}-DEC_{SU}). The abbreviations of the model parameters correspond to the labels in Fig. 3. For better comparability, the effective populations sizes of the structured meta-populations on Borneo and north of Lake Toba are given as the total effective sizes according to the formula $N_e = D \times N \times (1 + (1/(4 \times N \times m)))$, with D corresponding to the number of subpopulations, N to the mean subpopulation size and m to the total migration rate per individual per generation within the meta-population (Nichols *et al.* 2001).

Warren *et al.* 2001; Zhang *et al.* 2001; Steiper 2006; Nater *et al.* 2011).

The discrepancy between model-based species split estimates using exclusively autosomal data and mtDNA divergence time estimates from phylogenetic methods is owed to two idiosyncrasies in the biology of orang-utans. First, due to the pronounced philopatric tendencies of female orang-utans (Galdikas 1995; Arora *et al.* 2012; Nietlisbach *et al.* 2012; van Noordwijk *et al.* 2012), mtDNA has likely experienced only little if any gene flow between the two species after the species split. Therefore, the coalescent time of island-specific mitochondrial lineages is expected to predate the population split between Borneo and Sumatra, depending on N_e in the ancestral population (Nichols 2001). Second, due to male-mediated gene flow, model-based approaches using solely autosomal data are likely to underestimate species split times, as disentangling the contributions of migration and split time remains challenging (Hey & Nielsen 2004). The recent split time estimates from autosomal genomic data might reflect the end of an initial period of frequent, but strictly male-driven gene flow after the species split. Such complex temporal fluctuations in migration rates, as expected during glacial cycles for Sundaland species, are so far not properly addressed in any applied demographic model. Still, combining markers with different inheritance patterns as carried out in this study is likely to improve the estimates of both migration rates and split times in species with sex-biased dispersal such as orang-utans.

Our findings of recent gene flow between Bornean and Sumatran orang-utans are in agreement with previous observations (Muir *et al.* 2000; Verschoor *et al.* 2004; Becquet & Przeworski 2007). In their genomic study,

Locke *et al.* (2011) found an unexpectedly high incidence of low-frequency mutations shared between Borneo and Sumatra, which also hints at recent gene flow between the two islands. Contrary to studies indicating the presence of impassable dispersal barriers on the exposed Sunda shelf, due to either large river systems (Harrison *et al.* 2006) or a putative savannah corridor (Gathorne-Hardy *et al.* 2002; Bird *et al.* 2005), it seems that habitat conditions during glacial periods did at least sporadically allow male orang-utans to cross the exposed Sunda shelf. However, given the strict and long-lasting separation of mtDNA lineages on both islands (Nater *et al.* 2011), it appears that the exposed shelf was not covered with forest able to sustain orang-utan populations over prolonged periods. In fact, large parts of the Sunda shelf between Borneo and Sumatra were covered with nutrient-poor sandy soils (Bird *et al.* 2005; Slik *et al.* 2011). Forests on such soil types are characterized by low growth and productivity (Paoli *et al.* 2010). These constraints might explain why orang-utan populations on both islands could not expand onto the exposed shelf to an extent where population admixture and thus exchange of mtDNA lineages was possible.

Glacial cycles and population size changes

As we also tested models that incorporated sudden population size changes, we were able to detect signals of a population bottleneck on Borneo. In contrast to Sumatra, the currently observed pattern of strong population differentiation on Borneo (Warren *et al.* 2001; Arora *et al.* 2010) seems to have been established only recently, as parameter estimation indicated that Bornean orang-utans were organized at least temporarily as a single panmictic population before ~80 ka. At ~140 ka, the ancient population on Borneo experienced a sudden drop in N_e from ~17 000 to ~2500 individuals, which then recovered again to the current total N_e of ~6000 for all Bornean orang-utans. Such a change in both N_e and population structure could be explained by a common Bornean refugium during either the penultimate (190–130 ka) or last (110–18 ka) glacial period, when the drier and more seasonal climate might have caused a drastic reduction of rainforest coverage on Borneo (Morley 2000; Gathorne-Hardy *et al.* 2002; Bird *et al.* 2005). Population contractions with subsequent expansions likely occurred multiple times on Borneo during Pleistocene glacial and interglacial cycles, but incorporating such complex population dynamics into a demographic model is currently not feasible with the data at hand.

Interestingly, a similar signal of a glacial refugium with subsequent population structuring, as observed in Bornean orang-utans, has been found in western gorillas (*Gorilla gorilla*). Using a demographic modelling approach comparable to our study, Thalmann *et al.* (2011) found that the two subspecies of western gorillas (*G. g. gorilla* and *G. g. diehli*) diverged only about ~18 ka, thus directly following the last glacial maximum (LGM) 19–26 ka (Clark *et al.* 2009). Furthermore, the ancient population of western gorillas exhibited a N_e of just ~2500 individuals as compared to 22 000 and 17 000 individuals in the two subspecies after the population split. Therefore, it seems that western gorillas, similar to Bornean orang-utans, were constrained to a relatively small refugial population during glacial periods from which they subsequently expanded when the climate got warmer and wetter during interglacials.

Geological processes and population size changes

Linking bottleneck signals to specific environmental processes is difficult due to the large confidence intervals associated with most parameter estimates. For instance, the 90% highest posterior density interval for the estimate of the start of the bottleneck on Borneo (21–348 ka) also overlaps with the Toba supereruption on northern Sumatra ~73 ka (Chesner *et al.* 1991). It has been hypothesized that this colossal explosive eruption might have had a strong global impact, causing a severe bottleneck in humans (Rampino & Ambrose 2000). However, evidence presented here points towards climatic changes during the glacial periods rather than the Toba supereruption as being the main cause for the detected bottleneck on Borneo, as our results showed that the supereruption did not even have a strong impact on the Sumatran populations despite their much closer geographic proximity. Models incorporating a severe bottleneck in the Sumatran populations around the time of the supereruption were clearly rejected, and the signal of a recent population decline in Sumatra was considerably younger than the Toba supereruption. Studies indicate that the destruction caused by the Toba supereruption had been geographically limited, as shown by the distribution of rainforest refugia in South-East Asia (Gathorne-Hardy *et al.* 2002), including on Mentawai Island around 350 km from the Toba caldera (Gathorne-Hardy & Harcourt-Smith 2003), as well as the similar composition of South-East Asian fossil sites before and after the date of the supereruption (Louys 2007). Given the proximity of contemporary populations of Sumatran orang-utans to the Toba caldera and the strong dependency of orang-utans on intact rain forest habitat, they are undoubtedly one of the most striking examples illustrating the limited impact of the Toba supereruption on the local flora and fauna in South-East

Asia. However, the lack of bottleneck signals in the Sumatran populations does not imply that the activity of the Toba volcano did not influence the population history of Sumatran orang-utans at all. Rather, the results of this study, as well as previous findings (Nater *et al.* 2011, 2013), highlighted that the Toba eruptions must have repeatedly caused devastating damage to the local surroundings, which led to a long-lasting separation of gene pools north and south of Lake Toba.

In contrast to Toba as cause for the bottleneck on Borneo, a contraction of rainforests following the colder and drier climate during the last glacial period explains the absence of a similar bottleneck in the Sumatran population history well. During the generally drier glacial periods, large parts of Sumatra experienced considerably more rain fall compared to Borneo (Whitten *et al.* 2000; Gathorne-Hardy *et al.* 2002), because the Barisan mountain range running the length of Sumatra acted as a barrier for the wet monsoon winds, causing high precipitation along its western slopes (Whitten *et al.* 2000). This mountain ridge effect in combination with the close proximity to the sea during glacial periods, when sea levels were low, might have allowed large areas of rainforest to persist on Sumatra during glacial periods (Gathorne-Hardy *et al.* 2002). Thus, Sumatran orang-utans were almost certainly not forced into glacial refugia to the same extent as Borneans.

Anthropogenic impacts on orang-utan populations

While Sumatran orang-utans did not seem to go through glacial bottlenecks, we found evidence for recent and drastic declines in population sizes north and south of Lake Toba. These signals of population decline cannot be attributed to the large-scale human-induced habitat degradation that started in the last century (Rijksen & Meijaard 1999), of which genetic signals were found in previous studies of Bornean orang-utans (Goossens *et al.* 2006; Sharma *et al.* 2012). Rather, our results point towards an earlier decline in the Late Pleistocene or Early Holocene. In the Late Pleistocene, orang-utans went extinct on the South-East Asian mainland as well as in many Sundaland regions (Jablonski 1998; Rijksen & Meijaard 1999; Delgado & Van Schaik 2000). Furthermore, the Pleistocene–Holocene boundary is characterized by the disappearance of many large-bodied animals worldwide (Koch & Barnosky 2006), including large parts of the megafauna in South-East Asia (Louys *et al.* 2007). The increased occurrence of megafaunal extinctions during this period has been attributed to climatic changes following the LGM, the impact of human hunting and human-induced habitat changes, or the combination of these two factors (reviewed in Koch & Barnosky 2006).

Both climatic and anthropogenic factors might have played a role in the decline and local extinctions of orang-utan populations in the Late Pleistocene. During the LGM, the drier and more seasonal climate caused a shifting of zones of evergreen rainforest towards the equator (Flenley 1998; Jablonski 1998; Morley 2000), likely causing populations in southern China to go extinct. The warmer climate following the LGM was accompanied by rising sea levels, which drastically increased the extent of coastlines in Sundaland (Voris 2000). This enlargement of coastal habitat might have promoted an expansion of early modern humans on Sundaland, leading to increased hunting pressure on large-bodied animals, including orang-utans (Hill *et al.* 2007; Soares *et al.* 2008). Such hunting by modern humans might have caused the local extinctions of orang-utans on many Sundaland islands and led to a strong decline in Sumatran populations north and south of Lake Toba. Bornean orang-utans did not seem to be as strongly affected by human hunting, probably because the large size and low productivity of Borneo left enough inland areas with relatively low human densities (Delgado & Van Schaik 2000).

Our modelling approach revealed that the two recognized orang-utan species experienced drastically different demographic histories. Sumatran orang-utans exhibit a deep and temporally stable population structure, including an old divergence of gene pools north and south of Lake Toba with limited amount of gene flow over the Toba caldera. The populations on Sumatra went recently through a strong decline, which, in combination with strong population structure, explains the high genetic diversity found in recent genomic studies despite their low current census size (Locke *et al.* 2011; Prado-Martinez *et al.* 2013). In contrast, we find that the population structure currently observed within Bornean orang-utans has been established only recently and the population went through at least one bottleneck most likely associated with a glacial refugium.

These results strongly suggest that special consideration needs to be given to demographic factors when analysing adaptive evolutionary processes in great apes. Due to their strong dependence on intact forest habitat, most great ape taxa were severely affected by the climate shifts during glacial periods, which were accompanied by drastic changes in forest coverage in the tropics (Flenley 1998; Morley 2000). Accordingly, great ape populations experienced population bottlenecks, founder events, population expansions and population structuring as recent as 15 000 years ago (Clark *et al.* 2009). Given the long generation time of all great apes (18–30 years, Wich *et al.* 2009), great ape populations will likely not have reached an equilibrium state for most genomic regions. Thus, population expansions

and substructuring caused by relatively recent climatic changes might produce erroneous signals of selective sweeps if demography is not taken into account. Our results therefore emphasize the need to further advance the development of tools to jointly estimate demography and selection in order to unravel the convoluted evolutionary history of great apes (Li *et al.* 2012).

Acknowledgements

We are indebted to Pirmin Nietlisbach, Nicole Ponta, Livia Gerber, Corinne Ackermann and Kai Ansmann for providing valuable laboratory work for this study. We thank Erik Willems for the Sundaland map. Laurentius N. Ambu, Maria A. van Noordwijk, Helen Morrogh-Bernard, Cheryl Knott, Noko Kuze, Tomoko Kanamori, Joko Pamungkas, Dyah Perwitasari-Farajallah and Muhammad Agil provided orang-utan samples that were analysed for this study or helped with administration in Indonesia and Malaysia. Special thanks goes to Daniel Wegmann for his technical help with the ABCTOOLBOX software. We are grateful to three anonymous reviewers for their valuable comments. This project was financially supported by the Swiss National Science Foundation (Grant no. 3100A-116848 to MK and CPvS), Forschungskredit of the University of Zurich (Grant no. 57020601 to MPG), Messerli Foundation, A.H.-Schultz Foundation and Claraz Schenkung. Furthermore, we thank the following institutions for supporting our research: Primate Research Center of the Bogor Agricultural University (IPB), Indonesian State Ministry of Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), Sabah Wildlife Department, Taman National Gunung Leuser (TNGL), Borneo Orangutan Survival Foundation (BOSF), Leuser International Foundation (LIF) and Badan Pengelola Kawasan Ekosistem Leuser (BPKEL).

References

Arora N, Nater A, van Schaik CP *et al.* (2010) Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (*Pongo pygmaeus*). *Proceedings of the National Academy of Sciences*, **107**, 21376–21381.

Arora N, van Noordwijk MA, Ackermann C *et al.* (2012) Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (*Pongo pygmaeus*). *Molecular Ecology*, **21**, 3352–3362.

Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Bird MI, Taylor D, Hunt C (2005) Environments of insular Southeast Asia during the last glacial period: a savanna corridor in Sundaland? *Quaternary Science Reviews*, **24**, 2228–2242.

Boulesteix AL, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44.

Chesner CA, Rose WI, Deino A, Drake R, Westgate JA (1991) Eruptive history of Earths largest Quaternary caldera (Toba, Indonesia) clarified. *Geology*, **19**, 200–203.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, **186**, 983–995.

Clark PU, Dyke AS, Shakun JD *et al.* (2009) The last glacial maximum. *Science*, **325**, 710–714.

Cook SR, Gelman A, Rubin DB (2006) Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, **15**, 675–692.

Delgado RA, Van Schaik CP (2000) The behavioral ecology and conservation of the orangutan (*Pongo pygmaeus*): a tale of two islands. *Evolutionary Anthropology*, **9**, 201–218.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Excoffier L, Foll M (2011) FASTSIMCOAL: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, **27**, 1332–1334.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S (2006) Demographic history and genetic differentiation in apes. *Current Biology*, **16**, 1133–1138.

Flenley JR (1998) Tropical forests under the climates of the last 30,000 years. *Climatic Change*, **39**, 177–197.

Galdikas BMF (1995) Social and reproductive behavior of wild adolescent female orangutans. In: *The Neglected Ape* (eds Nadler RD, Galdikas BFM, Sheeran LK, Rosen N), pp. 163–182. Plenum Press, New York.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Gathorne-Hardy FJ, Harcourt-Smith WEH (2003) The supereruption of Toba, did it cause a human bottleneck? *Journal of Human Evolution*, **45**, 227–230.

Gathorne-Hardy FJ, Syaukani Davies RG, Eggleton P, Jones DT (2002) Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. *Biological Journal of the Linnean Society*, **75**, 453–466.

Goossens B, Chikhi L, Jalil MF *et al.* (2005) Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. *Molecular Ecology*, **14**, 441–456.

Goossens B, Chikhi L, Ancrenaz M *et al.* (2006) Genetic signature of anthropogenic population collapse in orang-utans. *PLos Biology*, **4**, 285–291.

Greminger MP, Stölting KN, Nater A *et al.* (2014) Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics*, **15**, 16. doi:10.1186/1471-2164-15-16.

Groves CP (2001) *Primate Taxonomy*. Smithsonian Institution Press, Washington, District of Columbia; London.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research*, **15**, 790–799.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, **62**, 255–265.

Harrison T, Krigbaum J, Manser J (2006) Primate biogeography and ecology on the Sunda shelf islands: a paleontological and zooarchaeological perspective. In: *Primate Biogeography* (eds Lehman SM, Fleagle JG), pp. 331–372. Springer, New York, USA.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hill C, Soares P, Mormina M *et al.* (2007) A mitochondrial stratigraphy for island southeast Asia. *American Journal of Human Genetics*, **80**, 29–43.

Jablonski NG (1998) The response of catarrhine primates to Pleistocene environmental fluctuations in East Asia. *Primates*, **39**, 29–37.

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics*, **27**, 155–156.

Kanthaswamy S, Kurushima JD, Smith DG (2006) Inferring *Pongo* conservation units: a perspective based on microsatellite and mitochondrial DNA analyses. *Primates*, **47**, 310–321.

Koch PL, Barnosky AD (2006) Late quaternary extinctions: state of the debate. *Annual Review of Ecology Evolution and Systematics*, **37**, 215–250.

Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics*, **184**, 243–252.

Li J, Li H, Jakobsson M *et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, **21**, 28–44.

Locke DP, Hillier LW, Warren WC *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.

Louys J (2007) Limited effect of the Quaternary's largest supereruption (Toba) on land mammals from Southeast Asia. *Quaternary Science Reviews*, **26**, 3108–3117.

Louys J, Curnoe D, Tong HW (2007) Characteristics of Pleistocene megafauna extinctions in Southeast Asia. *Palaeogeography Palaeoclimatology Palaeoecology*, **243**, 152–173.

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH (2011) Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genetics*, **7**, e1001319. doi:10.1371/journal.pgen.1001319.

Mailund T, Halager AE, Westergaard M *et al.* (2012) A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genetics*, **8**, e100312. doi:10.1371/journal.pgen.1003125.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15324–15328.

Mevik BH, Wehrens R (2007) The PLS package: principal component and partial least squares regression in R. *Journal of Statistical Software*, **18**, 1–28.

Morley RJ (2000) *Origin and Evolution of Tropical Rain Forests*. Wiley, Chichester.

Muir CC, Galdikas BMF, Beckenbach AT (2000) mtDNA sequence diversity of orangutans from the islands of Borneo and Sumatra. *Journal of Molecular Evolution*, **51**, 471–480.

Nater A, Nietlisbach P, Arora N *et al.* (2011) Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant orangutans (genus: *Pongo*). *Molecular Biology and Evolution*, **28**, 2275–2288.

Nater A, Arora N, Greminger MP *et al.* (2013) Marked population structure and recent migration in the critically endangered Sumatran orangutan (*Pongo abelii*). *Journal of Heredity*, **104**, 2–13.

Nei M, Suzuki Y, Nozawa M (2010) The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, **11**, 265–289.

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, **16**, 358–364.

Nichols RA, Bruford MW, Groombridge JJ (2001) Sustaining genetic variation in a small population: evidence from the Mauritius kestrel. *Molecular Ecology*, **10**, 593–602.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nietlisbach P, Arora N, Nater A *et al.* (2012) Heavily male-biased long-distance dispersal of orang-utans (genus: *Pongo*), as revealed by Y-chromosomal and mitochondrial genetic markers. *Molecular Ecology*, **21**, 3173–3186.

van Noordwijk MA, Arora N, Willems EP *et al.* (2012) Female philopatry and its social benefits among Bornean orangutans. *Behavioral Ecology and Sociobiology*, **66**, 823–834.

Paoli GD, Wells PL, Meijaard E *et al.* (2010) Biodiversity conservation in the REDD. *Carbon Balance and Managment*, **5**, 7.

Prado-Martinez J, Sudmant PH, Kidd JM *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rampino MR, Ambrose SH (2000) Volcanic winter in the Garden of Eden: the Toba supereruption and the late Pleistocene human population crash. In: *Volcanic Hazards and Disasters in Human Antiquity* (eds McCoy FW, Heiken G), pp. 71–82. Geological Society of America, Boulder.

Rice WR (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.

Rijksen HD, Meijaard E (1999) *Our Vanishing Relative: The Status of Wild Orang-Utans at the Close of the Twentieth Century*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Robert CP, Cornuet JM, Marin JM, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15112–15117.

Sharma R, Arora N, Goossens B *et al.* (2012) Effective population size dynamics and the demographic collapse of Bornean orang-utans. *PLoS ONE*, **7**, e49429. doi:10.1371/journal.pone.0049429.

Slik JWF, Aiba SI, Bastian M *et al.* (2011) Soils on exposed Sunda Shelf shaped biogeographic patterns in the equatorial forests of Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 12343–12347.

Soares P, Trejaut JA, Loo JH *et al.* (2008) Climate change and postglacial human dispersals in Southeast Asia. *Molecular Biology and Evolution*, **25**, 1209–1218.

Stadler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, **182**, 205–216.

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, **22**, 63–73.

Steiper ME (2006) Population history, biogeography, and taxonomy of orangutans (Genus: *Pongo*) based on a population genetic meta-analysis of multiple loci. *Journal of Human Evolution*, **50**, 509–522.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Thalmann O, Wegmann D, Spitzner M *et al.* (2011) Historical sampling reveals dramatic demographic changes in western gorilla populations. *BMC Evolutionary Biology*, **11**, 85. doi:10.1186/1471-2148-11-85.

Verschoor EJ, Langenhuijzen S, Bontjer I *et al.* (2004) The phylogeography of orangutan foamy viruses supports the theory of ancient repopulation of Sumatra. *Journal of Virology*, **78**, 12712–12716.

Voris HK (2000) Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *Journal of Biogeography*, **27**, 1153–1167.

Warren KS, Verschoor EJ, Langenhuijzen S *et al.* (2001) Speciation and intrasubspecific variation of Bornean orangutans, *Pongo pygmaeus pygmaeus*. *Molecular Biology and Evolution*, **18**, 472–480.

Wegmann D, Leuenberger C, Excoffier L (2009a) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Wegmann D, Leuenberger C, Excoffier L (2009b) *Using* ABCTOOLBOX. http://cmpg.unibe.ch/software/ABCtoolbox/ABCtoolbox_manual.pdf.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCTOOLBOX: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116. doi:10.1186/1471-2105-11-116.

Whitten T, Damanik SJ, Anwar J, Hisyam N (2000) *The Ecology of Sumatra*. Periplus Editions Ltd., Hong Kong.

Wich SA, Meijaard E, Marshall AJ *et al.* (2008) Distribution and conservation status of the orang-utan (*Pongo* spp.) on Borneo and Sumatra: how many remain? *Oryx*, **42**, 329–339.

Wich SA, deVries H, Ancrenaz M *et al.* (2009) Orangutan life history variation. In: *Orangutans: Geographic Variation in Behavioral Ecology and Conservation* (eds Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP), pp. 65–75. Oxford University Press, Oxford.

Williams MAJ, Ambrose SH, van der Kaars S *et al.* (2009) Environmental impact of the 73 ka Toba super-eruption in South Asia. *Palaeogeography Palaeoclimatology Palaeoecology*, **284**, 295–314.

Wilson IJ, Weale ME, Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society Series a-Statistics in Society*, **166**, 155–188.

Xu XF, Arnason U (1996) The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. *Journal of Molecular Evolution*, **43**, 431–437.

Zhang YW, Ryder OA, Zhang YP (2001) Genetic divergence of orangutan subspecies (Pongo pygmaeus). *Journal of Molecular Evolution*, **52**, 516–526.

Zhi L, Karesh WB, Janczewski DN *et al.* (1996) Genomic differentiation among natural populations of orang-utan (Pongo pygmaeus). *Current Biology*, **6**, 1326–1336.

A.N., M.P.G., C.P.vS. and M.K. designed the study; B.G., I.S., E.J.V. and K.S.W. provided samples; A.N., M.P.G. and N.A. performed laboratory procedures; A.N. and M.P.G. conducted genetic data analysis; A.N. performed demographic modelling; A.N. wrote the manuscript; M.P.G., N.A., C.P.vS. and M.K. critically revised the manuscript and provided comments at all stages; B.G., E.J.V. and K.S.W. edited the final manuscript.

Data accessibility

Sequence alignments, microsatellite genotypes, summary statistics, input files and custom-made scripts used in this study are available on Dryad: doi:10.5061/dryad.1jv55. Sequence data published previously are accessible under GenBank Accession nos HQ912716–HQ912752 (Table S2, Supporting information).

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Primers used for amplification and sequencing of four autosomal regions and one X-chromosomal region.

Table S2 List of sequence loci used in the ABC analysis.

Table S3 List of microsatellite loci used in the ABC analysis.

Table S4 Mutation rate estimates of sequence loci.

Table S5 Parameterisation and parameter prior distributions for all 2-population models.

Table S6 Parameterisation and parameter prior distributions for all 10-population models.

Table S7 Summary statistics used for Approximate Bayesian Computation.

Table S8 Model fits of all tested demographic models.

Table S9 Accuracy of different point estimators in parameter estimation.

Table S10 Accuracy of parameter estimation under different tolerance levels.

Fig. S1 Pr(Data | K) and deltaK statistics for all STRUCTURE runs.

Fig. S2 Structure plot for 25 microsatellite markers used for the demographic modelling.

Fig. S3 Gene trees based on sequence data of six different loci.

Fig. S4 Cross validation of the parameter estimation.

Fig. S5 First 16 principal components of the posterior predictive distribution for the selected model (IM10-BN_{BO}-DEC_{SU}).

Journal of Heredity 2013:104(1):2–13
doi:10.1093/jhered/ess065
Advance Access publication October 16, 2012

Marked Population Structure and Recent Migration in the Critically Endangered Sumatran Orangutan (*Pongo abelii*)

Alexander Nater, Natasha Arora, Maja P. Greminger, Carel P. van Schaik, Ian Singleton, Serge A. Wich, Gabriella Fredriksson, Dyah Perwitasari-Farajallah, Joko Pamungkas, and Michael Krützen

From the Anthropological Institute & Museum, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland (Nater, Arora, Greminger, van Schaik, and Krützen); Foundation for a Sustainable Ecosystem (YEL), Medan, Indonesia (Singleton and Fredriksson); Sumatran Orangutan Conservation Programme (PanEco-YEL), Medan, Indonesia (Singleton, Wich, and Fredriksson); Research Centre in Evolutionary Anthropology and Palaeoecology, Liverpool John Moores University, Liverpool, United Kingdom (Wich); Primate Research Center, Bogor Agricultural University, Bogor, Indonesia (Perwitasari-Farajallah and Pamungkas); Department of Biology, Bogor Agricultural University, Bogor, Indonesia (Perwitasari-Farajallah); Department of Animal Infectious Diseases and Veterinary Public Health, Bogor Agricultural University, Bogor, Indonesia (Pamungkas).

Address correspondence to Alexander Nater at the address above, or e-mail: a.nater@aim.uzh.ch.

Abstract

A multitude of factors influence how natural populations are genetically structured, including dispersal barriers, inhomogeneous habitats, and social organization. Such population subdivision is of special concern in endangered species, as it may lead to reduced adaptive potential and inbreeding in local subpopulations, thus increasing the risk of future extinctions. With only 6600 animals left in the wild, Sumatran orangutans (*Pongo abelii*) are among the most endangered, but also most enigmatic, great ape species. In order to infer the fine-scale population structure and connectivity of Sumatran orangutans, we analyzed the most comprehensive set of samples to date, including mitochondrial hyper-variable region I haplotypes for 123 individuals and genotypes of 27 autosomal microsatellite markers for 109 individuals. For both mitochondrial and autosomal markers, we found a pronounced population structure, caused by major rivers, mountain ridges, and the Toba caldera. We found that genetic diversity and corresponding long-term effective population size estimates vary strongly among sampling regions for mitochondrial DNA, but show remarkable similarity for autosomal markers, hinting at male-driven long-distance gene flow. In support of this, we identified several individuals that were most likely sired by males originating from other genetic clusters. Our results highlight the effect of natural barriers in shaping the genetic structure of great ape populations, but also point toward important dispersal corridors on northern Sumatra that allow for genetic exchange.

Key words: *conservation, gene flow, Great apes, microsatellites, Sundaland*

Most natural populations do not behave like single units, in which random mating occurs over the entire distribution (Kimura and Weiss 1964). Rather, most populations are genetically structured, the extent of which is determined by several factors. Geographical factors include both isolation by distance (Wright 1943) and physical barriers impeding gene flow across them, such as mountain ridges, rivers, and deserts. Ecological factors concern the distribution of resources and predators, which may lead to an aggregation of individuals within high-quality habitat patches (Slatkin 1987). A third category includes social, mating, and dispersal behaviors. Gregarious species, where individuals live in social groups, often show a marked population structure even in the complete absence of obvious geographical or ecological factors (Storz 1999; Ross 2001). Yet, strong genetic structuring imposed by limited dispersal has also been found in non-gregarious species. This is because in both gregarious and non-gregarious species it is potentially advantageous for individuals to show some degree of philopatry, as in the natal area food resources are familiar and kin is available

for social support (Johnson and Gaines 1990; Handley and Perrin 2007). Moreover, dispersal is usually heavily biased toward one sex, because one major benefit of dispersal is the avoidance of inbreeding (Bengtsson 1978; Pusey 1987). As a consequence, the extent of observed genetic structure may vary greatly depending on the inheritance mode of the genetic marker system used to investigate such patterns.

The underlying genetic structure of populations is especially important from a conservation perspective. Genetic structure may lead to local isolation of gene pools, resulting in effective subpopulation sizes that are only a fraction of the effective population size in a population without substructure (Charlesworth 2009). This has three important evolutionary consequences. First, lower effective sizes of subpopulations lead to stronger genetic drift effects and a reduced number of mutation events in each subpopulation. As a consequence, genetic diversity within each subpopulation will be lower compared with that of an unstructured population. Moreover, deleterious mutations that would be eliminated by background selection in unstructured populations might become fixed in small subpopulations, thus reducing the average population fitness (Hedrick and Kalinowski 2000; Reed and Frankham 2003). Second, population structuring increases the chance of mating among relatives, therefore causing potential loss of fitness due to inbreeding depression (Hedrick and Kalinowski 2000). Third, local separation of genetic variants will allow different selection pressures to act on specific subpopulations, thus allowing for adaptations to specific local environmental conditions (Williams 1966; Kawecki and Ebert 2004). While local adaptations raise the average fitness of subpopulations in a constant environment, the loss of genetic diversity reduces the potential of the subpopulations to adapt to changing environmental conditions and therefore carries greater risks of future extinctions (Reed and Frankham 2003). All these negative effects, however, can be counterbalanced by gene flow among subpopulations (Slatkin 1987). Therefore, knowledge about the extent to which genetic diversity is structured and exchanged across the range of a species is crucial to predict the long-term survival of populations and to implement effective conservation measures.

Population subdivision is a major concern in large-bodied animals with small population sizes, slow life histories, and low rates of reproduction, as such taxa are especially vulnerable to the aforementioned negative effects of population fragmentation (Hedrick and Kalinowski 2000). Great apes are of special interest in investigating the causes and consequences of population subdivision, not only because studying their population histories can reveal valuable insights into the evolution of modern humans, but also because all extant species are listed as endangered or even critically endangered (IUCN 2011). Furthermore, great apes show variation in dispersal patterns, which affects the genetic structuring of populations. For instance, chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) show female-biased dispersal (Tautz et al. 1999; Mitani et al. 2002), whereas males are the dispersing sex in orangutans (*Pongo* spp.) (Singleton and van Schaik 2002; Morrogh-Bernard et al. 2011; van Noordwijk et al. 2012; Arora et al. 2012), as is the case in most mammals (Dobson 1982). In contrast, in gorillas (*Gorilla* spp.), both sexes disperse, even though mean dispersal distance is different between males and females (Douadi et al. 2007).

In the past, a substantial body of work has investigated population structure in great apes, such as in chimpanzees (Becquet et al. 2007; Gonder et al. 2011), bonobos (Eriksson et al. 2004; Eriksson et al. 2006), gorillas (Bergl and Vigilant 2007; Guschanski et al. 2008), and Bornean orangutans (*Pongo pygmaeus*) (Warren et al. 2001; Goossens et al. 2005; Jalil et al. 2008; Arora et al. 2010). Yet, a detailed population genetic analysis of Sumatran orangutans (*Pongo abelii*) is still lacking, even though Sumatran orangutans are critically endangered (IUCN 2011). As of today, only an estimated 6600 individuals remain in the wild, when compared with about 54 000 Bornean orangutans (Wich et al. 2008). In contrast to the Bornean species, where three subspecies have been defined based on morphological characters (Groves 2001), no subspecies have been proposed for Sumatran orangutans.

Historically, Sumatran orangutans populated most of the Indonesian island of Sumatra, as evidenced by fossil finds and historical records (Rijksen and Meijaard 1999; Delgado and Van Schaik 2000). The current distribution is, however, restricted to small forest patches on the northern tip of Sumatra (Wich et al. 2008). Ecological and anthropogenic factors, such as prehistoric hunting and recent deforestation, have been suggested as explanations for the drastic range collapse of orangutans (Delgado and Van Schaik 2000). The comparatively limited range of Sumatran orangutans that remains today is subdivided by major rivers and mountain ridges. Moreover, the massive forest exploitation that started in the last century (Rijksen and Meijaard 1999) has caused severe habitat fragmentation, leaving habitat blocks of continuous forest that often harbor only a few hundred individuals (Wich et al. 2008). This habitat fragmentation in combination with the potentially very strong reproductive skew in Sumatran orangutan males (Setia and van Schaik 2007; Utami Atmoko et al. 2009) might have drastically reduced the effective sizes of local subpopulations, thus minimizing genetic diversity and posing a severe threat of future extinctions.

Sumatran orangutans show the strictest arboreality among all great apes (Delgado and Van Schaik 2000) and occur in two different rain-forest habitat types. Low-altitude peat-swamp forests offer high and constant food supplies and support the highest population densities (Husson et al. 2009). At lower densities, permanent populations of Sumatran orangutans can be found in dry-land forests up to an altitude of 1500 m above sea level or more (Wich et al. 2004; Husson et al. 2009). However, in non-riverine dry-land forests, the mast fruiting phenomenon causes extreme temporal fluctuations in food availability (Knott 1998; Husson et al. 2009), which may act as a strong selective pressure for adaptive traits related to prolonged food scarcity. Unfortunately, due to the absence of long-term field studies covering the entire extant range of Sumatran orangutans, little is known about variation in behavior, physiology, and morphology within this species that could hint at the presence of habitat specific adaptations.

The current lack of knowledge about the genetic structure of Sumatran orangutans is mainly caused by difficulties in obtaining samples with reliable provenance throughout the entire species' range. This factor prevented most previous genetic studies from interpreting the extraordinary high diversity on the mitochondrial DNA (mtDNA) level they found in Sumatran orangutans when compared with their Bornean sister species (Muir et al. 2000; Kanthaswamy et al. 2006; Steiper 2006). However, using samples with a well-defined geographic origin, Nater et al. (2011) showed that mitochondrial variation is strongly geographically structured on Sumatra. This study identified four distinct mitochondrial clusters in Sumatran orangutans, with divergence times of up to 3.5 million years. Similar, albeit less-pronounced patterns of geographical structuring of mtDNA was found in Bornean orangutans (Warren et al. 2001; Arora et al. 2010). However, mtDNA is not a good indicator of population structure and gene flow in species that show a strong male-bias in dispersal, like orangutans (Galdikas 1995; Singleton and van Schaik 2002; Morrogh-Bernard et al. 2011). In fact, using Y-chromosomal markers, Nater et al. (2011) showed that the deep divergence and strong geographic clustering observed with mtDNA is not present in the male population history, indicating long-distance migration by males across Sumatra. The amount of gene flow and the resulting extent of homogenization of autosomal gene pools among local subpopulations is, however, impossible to measure using only sex-linked marker systems.

In this study, we aimed to unravel patterns of genetic diversity and differentiation in Sumatran orangutans, using a combination of mitochondrial and autosomal genetic markers. We investigated the role of geographical, ecological, and behavioral factors underlying the fine-scale population structure and tested for connectivity among subpopulations. To achieve this, we analyzed the most comprehensive and largest set of orangutan samples from Sumatra to date, using samples from wild individuals originating from the entire species' range.

Materials and Methods

Sample Collection

Three different kinds of orangutan samples were analyzed for this study: First, fecal samples were collected non-invasively at long-term study sites. Second, in areas where animals were not habituated, we collected hair samples from deserted nests. Third, we obtained blood and hair samples of confiscated wild-born orangutans from the quarantine station of the Sumatran Orangutan Conservation Program (SOCP) in Medan, North Sumatra.

We obtained orangutan samples from seven different sampling regions (Figure 1A): Tripa (TR), North Aceh (NA, north of Tamiang River), West Leuser (WL), Central Leuser (CL, west side of Alas River), Langkat (LK, east of Alas River, south of Tamiang River), Batu Ardan (BA, east of Alas River, west of Lake Toba), and Batang Toru (BT, south of Lake Toba) (see Supplementary Table S1 online). Fecal and hair samples were collected and stored following the genetic sampling protocol of the orangutan network (http://www.aim.uzh.ch/orangutannetwork, last accessed August 24, 2012). All blood samples were taken during routine veterinary examination in the SOCP quarantine station. Blood samples were collected in standard EDTA blood collection tubes and stored at −20 °C.

The amount and reliability of information about the wild origin of rehabilitant orangutans varied considerably. We classified the provenance of these individuals as reliable if the location of confiscation was known in detail and if this location was near an extant wild orangutan population. The samples from rehabilitant orangutans that did not meet these criteria were classified as having unknown provenance and excluded from certain analyses (see below).

The collection and transport of samples was carried out in compliance with Indonesian and international regulations. Samples were exported from Indonesia to Zurich under the Convention on International Trade in Endangered Species (CITES permits 09717/IV/SATS-LN/2010, 07279/IV/SATS-LN/2009, 00961/IV/SATS-LN/2007, 06968/IV/SATS-LN/2005).

Laboratory Procedures

DNA from fecal, hair, and blood samples was extracted and processed following the procedures described in Nater et al. (2011). We used a set of 12 human-derived (Goossens et al. 2005) and 15 species-specific microsatellite markers (Nietlisbach et al. 2010) to genotype the orangutan samples. In order to minimize genotyping errors due to allelic dropout, we followed the real-time PCR approach from Morin et al. (2001), performing between two and seven independent PCR repetitions per sample. PCR conditions and fragment length analysis are described in Arora et al. (2010) and Nietlisbach et al. (2010). We were able to genotype 112 out of 162 samples for at least 24 microsatellite loci. The identity check revealed three and two samples that were present as a triplicate and a duplicate, respectively, resulting in 109 unique genotypes.

For the sequencing of the hyper-variable region I (HVRI) of the mtDNA d-loop, we used the same primers, PCR conditions, and sequencing chemistry as Arora et al. (2010), resulting in a final alignment of 457 base pairs. Some sequences were from samples with insufficient DNA quantity for successful microsatellite genotyping. To avoid duplicates in the HVRI dataset, we only included sequences from individuals that had either a distinct genotype or were sampled more than 50 km apart from other samples in the dataset, resulting in 123 HVRI sequences. The sequences are deposited on GenBank under the accession numbers JQ962945–JQ962972.

HVRI Median-Joining Network

A median-joining network (Bandelt et al. 1999) using all HVRI sequences was drawn using NETWORK v4.6.0.0 and NETWORK PUBLISHER v1.3.0.0 (http://www.fluxus-engineering.com, last accessed August 24, 2012). An epsilon value of zero and equal weighting of all nucleotide positions

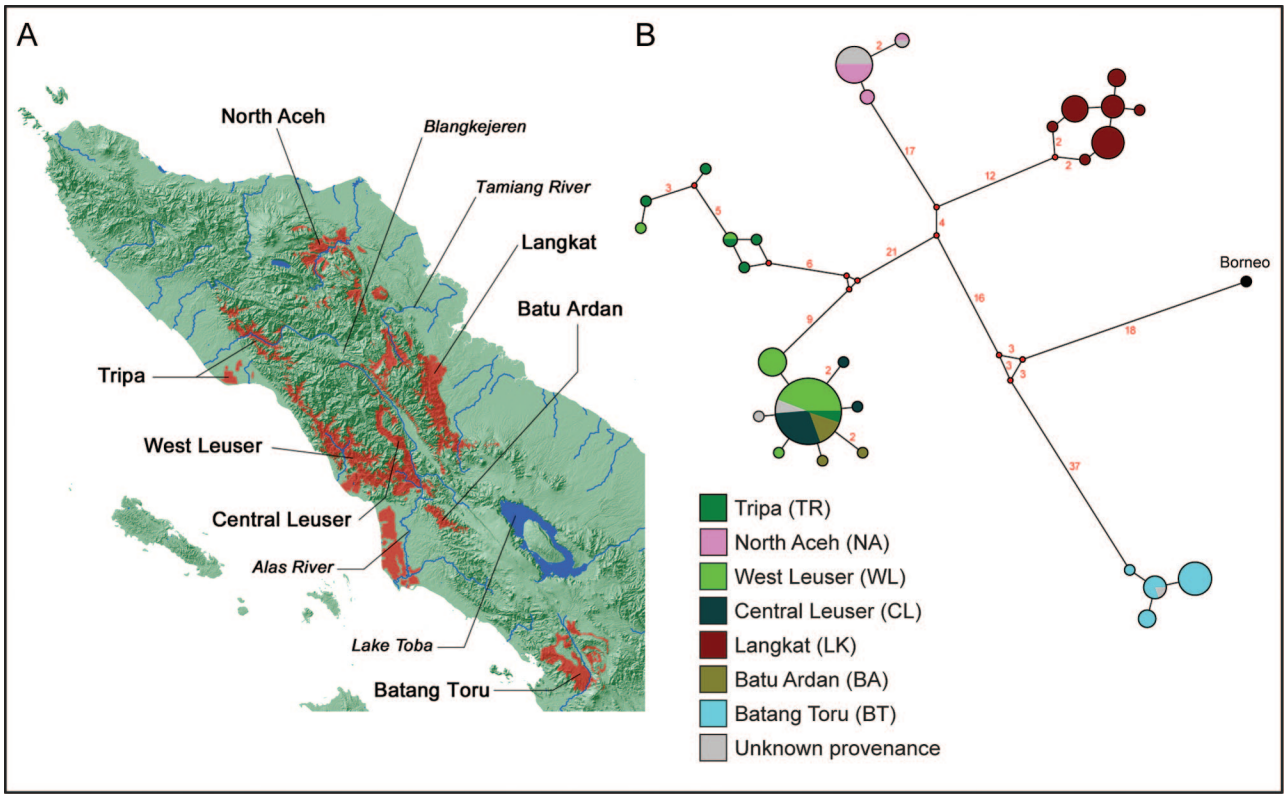


Figure 1. (**A**) Map of sampling regions in northern Sumatra. Labels in italics denote important geographic features. The red shading represents the current distribution of Sumatran orangutans. (**B**) Median-joining network of mitochondrial HVRI haplotypes. The red numbers in between the nodes indicate the number of mutational steps in between haplotypes (one step if not indicated otherwise). The size of each node is proportional to the number of individuals with the same haplotype.

was used for the network presented here. Using higher epsilon values or differently weighted transitions/transversions did not change the basic structure of the network.

Summary Statistics

We computed summary statistics and genetic differentiation measures for HVRI sequences and autosomal microsatellites using ARLEQUIN v3.5.1.2 (Excoffier and Lischer 2010). For both mitochondrial and autosomal datasets, we incorporated only samples with reliable provenance information. Based on this information, we divided the sample set *a priori* into seven sampling regions (Table 1).

To assess pairwise population differentiation, we calculated the differentiation measures Φ_{ST} (HVRI, Excoffier et al. 1992) and R_{ST} (microsatellites, Slatkin 1995). We used the Tamura and Nei distance correction (Tamura and Nei 1993) with a gamma value of 0.219 for the calculation of the genetic distance matrix for Φ_{ST}, as determined by the model selection test with jMODELTEST v0.1.1 (Posada 2008).

To infer the long-term effective population size N_e of the seven sampling regions, we calculated the estimators θ_π (based on the mean pairwise genetic distance between sequences; Tajima 1983) and θ_H (based on the heterozygosity of microsatellites; Ohta and Kimura 1973). Additionally, we used a likelihood-based estimator of θ (referred to as θ_L) using the software LAMARC v2.1.6 (Kuhner 2006). We applied the GTR+I nucleotide substitution model (Lanave et al. 1984) for the HVRI sequence data, which is the best-fitting of the supported models inferred by jMODELTEST, and the stepwise mutation model for the microsatellite data. The analysis was performed for each sampling region separately, and we used the Bayesian sampler with two chains of 1 000 000 steps each, sampling every 20th step and discarding the first 5000 samples as burn-in. The prior distribution of θ ranged from 10^{-5} to 10 (uniform on a natural logarithmic scale) and the starting value of θ was set to 0.01.

The different estimators of θ were used to calculate N_e, with θ equaling $N_e\mu$ for mitochondrial and $4N_e\mu$ for autosomal markers. Thus, these estimators allow inferring long-term N_e from a single population sample if the mutation rate is known. We used a mutation rate of 4.108×10^{-6} per site per generation for HVRI (Soares et al. 2009), assuming a generation time of 25 years (Wich et al. 2009), or 1×10^{-4} per locus per generation for the autosomal microsatellites (Schlötterer 2000).

Autosomal Genetic Structure

To assess genetic structure based on autosomal microsatellites, we first performed a principal component analysis

Table 1 Summary statistics for all examined orangutan sampling regions

		HVRI				Autosomal microsatellites				
Sampling region	Habitat[a]	$N_{Samples}$	θ_π[b]	HD[c]	N_e[d]	$N_{Samples}$	H_E[e]	θ_H[f]	N_e[g]	Census[h]
Tripa (TR)[i]	PSF	7	12.78	0.95	6808	9	0.64	1.68	4197	~380
North Aceh (NA)	DF	10	0.79	0.51	389	10	0.61	1.60	4009	~350
West Leuser (WL)	PSF	28	3.78	0.54	2013	21	0.61	1.61	4023	~3000
Central Leuser (CL)	DF	14	0.44	0.27	237	15	0.59	1.56	3901	~1100
Langkat (LK)	DF	26	1.40	0.80	747	24	0.64	1.66	4162	~1050
Batu Ardan (BA)	DF	8	0.78	0.46	417	9	0.59	1.57	3929	~300
Batang Toru (BT)	DF	18	0.96	0.65	503	8	0.63	1.63	4087	~550

[a]Prevailing habitat type; PSF, peat-swamp forest; DF, dry-land forest (Husson et al. 2009).
[b]Estimate of $\theta = N_e\mu$ based on the mean pairwise corrected nucleotide distance.
[c]Haplotypic diversity (Nei 1987).
[d]Effective population size, based on a mutation rate of 1.643×10^{-7} per site per year and a generation time of 25 years.
[e]Mean expected heterozygosity.
[f]Estimate of $\theta = 4N_e\mu$ based on the mean expected heterozygosity.
[g]Effective population size, based on a mutation rate of 10^{-4} per locus per generation.
[h]Estimated census size (Wich et al. 2008).
[i]The sampling region of Tripa includes coastal and highland areas.

(PCA) using the covariance-standardized method as implemented in the software GENALEX v6.41. Next, we used the Bayesian clustering algorithm implemented in the software STRUCTURE v2.3.3 (Pritchard et al. 2000) to identify distinct genetic clusters in the dataset. Because both methods do not require making *a priori* assumptions about genetic structure, we were able to include samples with unknown provenance. For the STRUCTURE analysis, we used the admixture model with correlated allele frequencies, a burn-in length of 3×10^5 steps followed by 3×10^6 MCMC steps. We ran the analysis with K values ranging from 1 to 10. For each K we performed 10 independent runs and averaged the ln Pr(Data | K) statistic over all iterations. Since the Pr(Data | K) estimator has been shown to overestimate K, as it frequently plateaus at higher values than the true number of K (Evanno et al. 2005), we also calculated the delta K statistic (Evanno et al. 2005), which gives a conservative estimate of K.

Migrant Detection

To assess the level of subpopulation connectivity, we identified individuals in the dataset that were either direct migrants or first generation offspring of direct migrants and local individuals. To achieve this, we used two different methods. First, given the strong geographic clustering of mtDNA haplotypes (Nater et al. 2011), we checked the median-joining network for individuals with reliable provenance that clustered with samples from another geographic region in order to detect direct migrants. Second, we used a Bayesian approach to assign individual genotypes to different subpopulations as either local individuals, direct migrants or F_1 admixed individuals, as implemented in the software BAYESASS 1.3 (Wilson and Rannala 2003). For this, we pre-assigned the individuals to the three different clusters identified in the previous STRUCTURE analysis and ran the MCMC analysis two times independently for 2.4×10^7 steps each, including a burn-in of 4×10^6 steps, with sampling every 2000 steps.

Table 2 Pairwise population differentiation values for HVRI (Φ_{ST}, above diagonal) and autosomal microsatellites (R_{ST}, below diagonal)

Φ_{ST}/R_{ST}	TR	NA	WL	CL	LK	BA	BT
TR	-	0.89***	0.58***	0.70***	0.95***	0.61**	0.97***
NA	0.05*	-	0.94***	0.99***	0.98***	0.98***	0.99***
WL	0.02	0.06**	-	0.04	0.96***	0.01	0.98***
CL	0.04*	0.11***	0.02	-	0.99***	0.02	1.00***
LK	0.02	0.02	0.05***	0.05***	-	0.98***	0.99***
BA	0.05	0.07*	0.07***	0.08**	0.00	-	0.99***
BT	0.12**	0.17***	0.14***	0.10***	0.08***	0.12**	–

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Both runs combined resulted in a total of 20 000 assignments for each individual.

Results

HVRI Median-Joining Network

The median-joining network (Figure 1B) showed a strong structuring of mtDNA haplotypes into four geographically distinct clusters: (1) Batang Toru, (2) Langkat, (3) Tripa, West Leuser, Central Leuser and Batu Ardan (referred to as West Alas cluster), and (4) North Aceh. We did not observe any haplotype sharing among these four clusters in our dataset of individuals with reliable provenance information.

Summary Statistics

The division of mitochondrial haplotypes into four distinct clusters as apparent in the mtDNA network correlated well with the Φ_{ST} statistic of genetic differentiation, as all comparisons between different clusters were highly significant (Table 2, above diagonal). However, within the West Alas

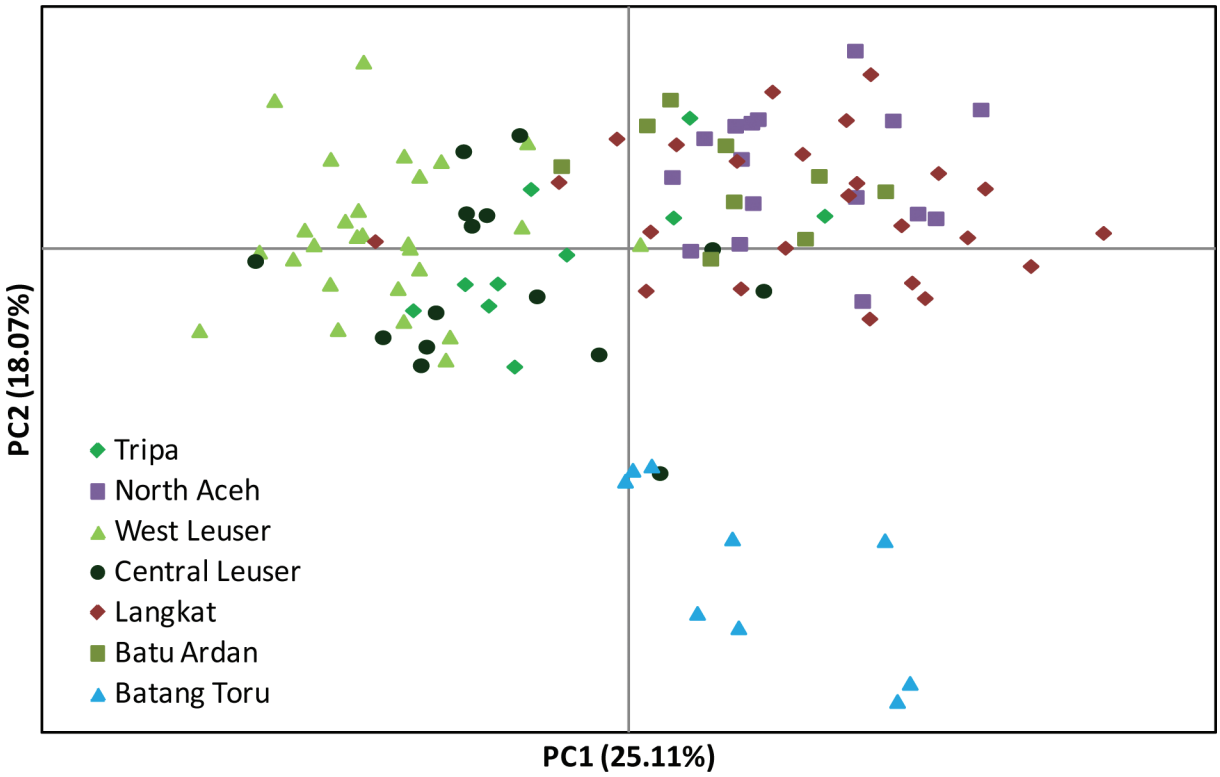


Figure 2. Principal component analysis of the autosomal microsatellite markers for all seven sampling regions.

cluster, the sampling region of Tripa was also significantly differentiated from all other regions in the same cluster. This differentiation points to highly different haplotype frequencies between Tripa and the other regions within this cluster, as these all share haplotypes among each other.

The R_{ST} measures for the microsatellites revealed additional information about the population structure beyond female philopatric patterns (Table 2, below diagonal). Three main patterns emerged. First, Batang Toru, the only sampling region south of Lake Toba, was highly differentiated from all other regions. Second, in contrast to high mtDNA differentiation, Tripa showed low R_{ST}-values to most other sampling regions, except Batang Toru. Third, the region of Langkat showed low differentiation to North Aceh, Tripa, and Batu Ardan.

The different estimators of θ revealed consistent patterns among the seven sampling regions, but estimates of θ for the microsatellite loci were consistently higher for θ_L when compared with θ_H (see Supplementary Table S2 online). We found that the genetic diversity estimates based on mtDNA and the corresponding N_e varied extensively across the different sampling regions (Table 1), as expected from the large differences in density estimates and habitat areas (Wich et al. 2008; Husson et al. 2009). In general, the estimated effective population sizes were similar to the census size estimates for most sampling regions (Wich et al. 2008). There was one striking exception. Tripa on the northwest coast exhibited the highest sequence diversity among the seven sampling regions and a N_e of nearly 7000 individuals, but contains among the smallest number of orangutans, with an estimated census size of less than 400 individuals. The Tripa region also showed a positive Tajima's *D* statistic and a multimodal pairwise mismatch distribution of HVRI sequences, indicating a recent population decline, while most other regions exhibited negative values of *D* and unimodal mismatch distributions, indicating recent expansions (Figure S3; see Supplementary Table S2 online). In contrast to the large regional variability for mtDNA, autosomal estimates of genetic diversity and N_e were remarkably similar among sampling regions (Table 1).

Autosomal Genetic Structure

The PCA revealed a geographically defined structure in the autosomal microsatellite data (Figure 2). The first principal component (PC) explained 25.11% of the total variance and distinguished between the sampling regions west and east of the Alas River. The region south of Lake Toba, Batang Toru, clusters with the regions east of the Alas River and cannot be distinguished with the first PC only. The second PC, explaining a further 18.07% of the variance, separated Batang Toru from all sampling regions north of Lake Toba. Therefore, by combining both PCs (explaining 43.18% of the total variance), there appears to be three clusters of sampling regions, separated from each other by the Alas River and Lake Toba. The separation was, however, not perfect, as the regions of WL, TR, BA, and CL showed outliers within the variation of other regions. The additional PCs did not seem to contain

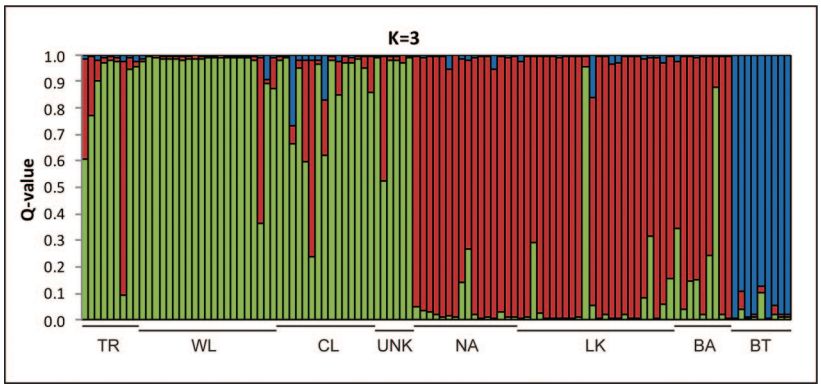


Figure 3. Results of the STRUCTURE analysis of the autosomal microsatellite markers for the most probable number of clusters ($K = 3$ according to delta K statistic). The membership coefficients Q shown are for the iteration with the highest likelihood. Samples are grouped by sampling region. The assignment is based on provenance record and mtDNA haplotype. UNK refers to samples with unknown provenance and ambiguous mtDNA assignment (belonging to the West Alas cluster).

any further information about geographic structuring of genotypes (Supplementary Figure S4).

The STRUCTURE analysis resulted in a clear signal for substructure in the Sumatran autosomal microsatellite dataset. Highest delta K was achieved for three clusters, while Pr(Data|K) peaked at five clusters (Supplementary Figure S5). At $K = 3$, the clusters corresponded largely to the mtDNA haplotype clusters described above, with some exceptions (Figure 3). First, the North Aceh and Langkat regions grouped together. Second, the region of Batu Ardan, which in the HVRI network assigned to the West Alas cluster, showed for autosomal markers a clear affinity to the Langkat and North Aceh regions. Third, the separation between the two genetic clusters north of Lake Toba (West Alas and Langkat/North Aceh) was not as sharp as for the mtDNA, as regions close to the geographic boundaries of the two clusters revealed a number of individuals with admixed genotypes. In contrast, samples from south of Lake Toba (Batang Toru) showed much less signals of admixture. Patterns of genetic admixture were also evident when the membership coefficients Q for each cluster were plotted in ranked order for all individuals for each cluster (Supplementary Figure S6). While all three curves showed two asymptotes at $Q = 0$ and $Q = 1$, multiple samples had Q-values between 0.2 and 0.8 (13 for West Alas, 13 for Langkat/North Aceh, and one for Batang Toru), indicating admixed ancestry.

A higher number of K did not result in a better resolution of sampling regions (Supplementary Figure S7). Since STRUCTURE often only identifies the uppermost level of hierarchical genetic structure (Evanno et al. 2005), we repeated the analysis for each of the three geographically defined clusters separately, using only samples that showed a membership coefficient of higher than 0.6 for a certain cluster in the first STRUCTURE analysis. None of the three clusters showed any sign of further substructure, as $K = 1$ returned the highest Pr(Data|K) values for all three clusters.

To test if part of the partitioning of the mitochondrial or autosomal genetic diversity can be explained by habitat type, we performed an AMOVA analysis with ARLEQUIN, where we divided the dataset into two groups corresponding to habitat type (peat-swamp forest versus dry-land forest, see Table 1). We included only samples from the West Alas cluster, as this is the only autosomal cluster that contains both habitat types. For autosomal microsatellites, habitat differences explain only 0.22% of the total variance, while over 97% is found within sampling regions (Table 3). For the mtDNA diversity, the variance component between habitat types is negative, indicating complete absence of any partitioning of genetic variance between habitat types.

Migrant Identification

All individuals showed congruence between their provenance record and their assigned mtDNA cluster. We did, however, identify three females and two males with high Q-values (>0.6) for a cluster that did not match their mtDNA haplotypes and provenance ($K = 3$, Figure 3). These individuals are unlikely to be direct migrants from the autosomal cluster they were assigned to in the STRUCTURE analysis. Rather, their natal range is indicated by their mtDNA haplotype, given that female orangutans have been shown to exhibit strong philopatric tendencies.

The BAYESASS analysis assigned migrant status to three of the five individuals previously identified in the STRUCTURE analysis as admixed or assigned to a cluster that did not match their mtDNA haplotype. In total, we found five individuals which have a less than 50% probability of being local in the cluster defined by their mtDNA haplotypes (Table 4). Only in one case, however, could we identify an admixed individual with significant statistical support ($P < 0.05$ of being local). This individual was a female with reliable provenance information, originating from the upper Alas valley in the Langkat region and carrying an mtDNA haplotype from the Langkat cluster. Her genotype, however,

Table 3 AMOVA of mitochondrial and autosomal microsatellite data between peat-swamp and dry-land forests within the West Alas cluster

	mtDNA		Autosomal microsatellites	
	Variance[a]	% Variance	Variance[a]	% Variance
Between habitat types	−0.74	−24.55	0.19	0.22
Among sampling regions, within habitat types	1.94*	64.71	1.87	2.17
Within sampling regions	1.80*	59.83	84.34*	97.61

*$P < 0.05$

[a]negative variance components indicate lack of genetic structure.

Table 4 List of individuals that show a probability of less than 0.5 to originate from the sampling cluster

					BAYESASS[d]		
Sample number	Sampling region[a]	Sex	mtDNA[b]	Q-value[c]	Local	Direct migrant	Admixed
BA2	BA (LK+NA)	Female	WA	0.876 (WA)	0.088 (LK+NA)	0.359 (WA)	0.553 (WA)
LK3	LK (LK+NA)	Female	LK	0.702 (LK+NA)	0.494 (LK+NA)	0.010 (WA)	0.496 (WA)
LK27	LK (LK+NA)	Female	LK	0.955 (WA)	0.004 (LK+NA)	0.365 (WA)	0.632 (WA)
LK7	LK (LK+NA)	Male	LK	0.673 (LK+NA)	0.409 (LK+NA)	0.002 (WA)	0.589 (WA)
TR4	TR (WA)	Male	WA	0.884 (LK+NA)	0.443 (WA)	0.228 (LK+NA)	0.329 (LK+NA)

[a]The autosomal genetic cluster to which most of the samples from the listed sampling regions assign is written in parentheses: WA, West Alas cluster, LK+NA, Langkat/North Aceh cluster, BT, Batang Toru cluster.
[b]mtDNA cluster assignment.
[c]Highest Q-value in the STRUCTURE analysis with $K = 3$.
[d]Posterior probabilities of the three classes in the BAYESASS analysis.

had a high membership coefficient to the West Alas cluster ($Q = 0.955$).

Discussion

Our study is the first to precisely locate and describe the geographic structuring of genetic diversity on mitochondrial and autosomal levels across the whole range of Sumatran orangutans. We were able to quantify the genetic diversity present within each of the seven sampling regions by analyzing the highly polymorphic HVRI region of the mtDNA and used that information to infer long-term effective population sizes of each sampling region. These estimates correlate strongly with recent census size estimates for most regions (Wich et al. 2008). Not surprisingly, the highest effective population sizes were observed for peat-swamp forests on the west coast of northern Sumatra, which also have the highest population density estimates (Husson et al. 2009). In one region, however, N_e and census size were in stark contrast to each other: the area of Tripa showed extraordinary high mitochondrial HVRI diversity and corresponding N_e in a comparatively small geographic region, which contains only an estimated 380 individuals. This signal points to a massive recent decline in the subpopulation size, which might have been caused by the dramatic and on-going habitat degradation in this area (van Schaik et al. 2001; Gaveau et al. 2009). It is plausible to assume that the lowland area along the northwest coast of Aceh was once completely covered with continuous peat-swamp forest and harbored thousands of orangutans (Gaveau et al. 2009). After decades of deforestation, current estimates indicate that all forests in the Tripa region will be irrecoverably lost by 2015–16 if forest destruction/conversion will continue at its current rate (Tata et al. 2010; Wich et al. 2011). There are other prominent examples in the literature highlighting discrepancies between large long-term N_e and small census sizes, which are linked to anthropogenic pressures. For example, heavy exploitation of gray (*Eschrichtius robustus*) and humpback whale (*Megaptera novaeangliae*) stocks due to whaling has led to dramatic population declines not reflected by long-term N_e (Roman and Palumbi 2003; Alter et al. 2007).

In contrast to the varying HVRI diversity found within different regions across the island, we obtained very homogenous genetic diversity estimates among sampling regions for autosomal microsatellite markers, resulting in N_e estimates of around 4000 or 10 000 individuals for each of the seven regions, depending on the estimator of θ. This striking discrepancy when compared with the HVRI estimates is most likely caused by pronounced male-biased dispersal and strong female philopatric tendencies in orangutans (Galdikas 1995; Singleton and van Schaik 2002; Morrogh-Bernard et al. 2011; Nietlisbach et al. 2012; van Noordwijk et al. 2012; Arora et al. 2012). Field studies have shown that female orangutans preferentially establish their home range overlapping with the home ranges of their maternal kin (Singleton and van Schaik 2002; van Noordwijk et al. 2012). Thus, mitochondrial DNA does get hardly, if at all, exchanged among neighboring geographic regions, and mtDNA diversity well reflects the number of orangutans in

the different local subpopulations. Males, in contrast, leave their natal area, a pattern linked to inbreeding avoidance (Pusey and Wolf 1996). Intense male-male competition (Utami Atmoko et al. 2009) may force young males to cover large distances before being able to settle down (Nietlisbach et al. 2012). Such widely dispersing males might distribute newly arisen alleles in the whole meta-population and recover alleles that have been lost locally due to genetic drift, thereby homogenizing the allele frequencies of autosomal markers among sampling regions. Thus, the highly similar levels of autosomal diversity in contrast to the large differences in mtDNA diversity across the island are a clear indicator of considerable male-mediated gene flow among these regions. The panmictic distribution of Y-haplotypes on Sumatra (Nater et al. 2011) provides further evidence for this male-driven homogenization of the gene pool.

Due to the use of multiple independent autosomal markers, we were able to investigate male-mediated gene flow in more detail. The cluster analysis with STRUCTURE showed that the strength of male-driven gene flow is not sufficient to completely homogenize allele frequencies among sampling regions, thus resulting in a clear pattern of geographically structured autosomal variation. The three clusters identified in the autosomal dataset were defined by geographical features. It appears that eruptions of the Toba volcano (Chesner et al. 1991) isolated the orangutans from Batang Toru, the region south of it, from the rest of the species occurring north of it. The high pairwise R_{ST}-values across Lake Toba provide further evidence of strong separating effects of the Toba eruptions, which have also led to a deep divergence of mtDNA haplotypes north and south of the caldera (Nater et al. 2011). The forests between these two areas might have been connected between major eruptions, but the combination of periodic separation and strong female philopatry has served to keep the populations from homogenizing. North of Lake Toba, the Alas River, part of the Barisan graben running the length of Sumatra (Verstappen 1973), divides the remaining regions into two distinct genetic clusters. The Alas valley was likely repeatedly blocked by volcanic material from the nearby Toba eruptions, turning the upper Alas river into a large lake for prolonged periods (van Schaik and Mirmanto 1985). This damming of the Alas River might have promoted the structuring of the gene pool north of Lake Toba. Interestingly, the habitat type does not seem to play a significant role in the structuring of autosomal diversity in Sumatran orangutans, indicating that dispersing males do not prefer to migrate to areas that ecologically resemble their natal habitat, and thus prevent more fine-tuned adaptation of orangutans to local habitat types.

Even though the STRUCTURE analysis revealed strong geographical structuring of the autosomal gene pool, we nevertheless found clear signals for recent gene flow across the island. First, the two sample regions of Langkat and North Aceh cannot be distinguished in the STRUCTURE analysis, even though these regions show a mitochondrial divergence of 0.85 Ma (Nater et al. 2011). Therefore, the observed low autosomal differentiation (R_{ST} = 0.02) points towards considerable levels of male-mediated gene flow after the two subpopulations were separated from each other. If this migratory contact with the Langkat region can be maintained, it will greatly help reducing inbreeding pressure on the small North Aceh subpopulation. As a second signal of gene flow, we found many admixed individuals in the STRUCTURE plot (Figure 3). Interestingly, these individuals were mostly sampled in regions close to the boundary of autosomal clusters, like Tripa, Central Leuser, and Langkat, supporting the idea of recent gene flow. Third, we were able to identify multiple individuals with substantial likelihoods of having paternal ancestry from another cluster. While only one individual shows good statistical support for being admixed ($P < 0.05$), it should be kept in mind that we sampled only an estimated 0.7–4.6% of all individuals per sampling region. Moreover, we only investigated migration among major autosomal clusters and not individual sampling regions, due to the impossibility to reliably discriminate them genetically.

Further investigation of the provenance of admixed individuals hinted toward an important corridor for gene flow between genetic clusters. Three of the five individuals identified as having admixed ancestry originate from the upper Alas valley near Blangkejeren, while a fourth admixed individual has been confiscated in the highlands of the Tripa area. These locations are all close to the area where the supposed boundaries of the West Alas, North Aceh, and Langkat clusters meet, and this highland area contains orangutan habitat with resident subpopulations. The presence of clear migration signals in this area underlines its critical importance as a connection among major subpopulations of Sumatran orangutans and therefore deserves special habitat conservation efforts.

Special consideration also needs to be given to the region of Batu Ardan, where there is a clear discrepancy between autosomal data and mtDNA structure, possibly due to male-mediated migration. This region, located between the Alas River and Lake Toba, shows a strong affinity of mtDNA haplotypes to the West Alas cluster, even though it is located on the opposite (eastern) side of the major Alas River. In fact, Batu Ardan shares a common haplotype with all regions on the western side, but also has two derived haplotypes that do not occur elsewhere. This supports the notion that the small Batu Ardan subpopulation could be the result of a recent colonization event from the western side of the Alas, probably due to a loop cut-off of the meandering river (Nater et al. 2011). However, for autosomal markers, we found that Batu Ardan reveals a high affinity to the adjacent Langkat/North Aceh cluster, from which it is separated by a deep river valley. This river might be passable by orangutans near its headwaters, allowing males to bring in autosomal alleles from the Langkat region. The notion that the recolonization from the west side of the Alas and subsequent influx of males from Langkat occurred after the forests recovered from the devastating Toba super-eruption around 73 kya (Chesner et al. 1991) is tempting but cannot yet be proven with the data at hand.

Sumatran orangutans are genetically deeply structured into at least three autosomally distinct clusters, despite regular male-mediated gene flow between the West Alas and the Langkat/North Aceh clusters, which occurred at least up to very recently and is probably still on-going. However, continuing habitat degradation is threatening the existence of orangutans on Sumatra in two ways. First, due to the shrinkage of suitable habitat area, the local subpopulation census sizes will be further reduced. Already today, only one of the three autosomal clusters, West Alas, harbors well over 1000 individuals. Second, through the destruction of important corridors for migration, genetic exchange with neighboring subpopulations will be disrupted. Both effects combined will inevitably lead to a substantial loss of genetic diversity with all its negative consequences (Reed and Frankham 2003). Especially the only remaining subpopulation south of Lake Toba, Batang Toru, is highly threatened in this regard. Given the genetic uniqueness of the orangutans in this area on both the mitochondrial and autosomal level and the fact that most of the forest in this area has no protected status (Wich et al. 2011), urgent measures are needed to preserve this indispensable reservoir of genetic diversity of Sumatran orangutans.

Orangutans are the least gregarious and the most arboreal of all great apes (Delgado and Van Schaik 2000). As such, comparing the observed patterns in Sumatran orangutans with those of other great ape species will aid the inference of factors underlying the observed population structure in these taxa. Previous genetic studies on great apes showed that rivers are one of the most important factors in shaping population structure and subspecies boundaries (e.g., gorillas: Anthony et al. 2007; Bornean orangutans: Goossens et al. 2005, Arora et al. 2010; chimpanzees: Becquet et al. 2007; bonobos: Eriksson et al. 2004). Our study supports these findings by identifying the Alas River as a major division line of genetic diversity within the range of Sumatran orangutans. Moreover, volcanic activities of the Toba region during the last 1.2 million years (Chesner et al. 1991) played another major role in the structuring of genetic diversity in Sumatran orangutans. Such a pattern of long-lasting isolation caused by volcanic activities has so far not been documented for great apes.

Given that Sumatran orangutans are critically endangered, knowledge of the extent to which human-induced habitat degradation is affecting the population structure is of critical importance for conservation efforts. Bergl and Vigilant (2007) revealed a pronounced substructure in the small Cross River gorilla population (*Gorilla gorilla diehli*) largely following the patterns of forest connectivity. Likewise, Goossens et al. (2005) showed that in Bornean orangutans, subpopulations in many of the isolated forest lots on the same side of the Kinabatangan River in Sabah, Malaysia, are significantly differentiated from each other, despite their close geographic proximity. Both studies highlight the adverse effects of anthropogenic forest degradation on the dispersal abilities of forest dwelling primates. Interestingly, we did not observe similar signals in Sumatran orangutans, despite their strict arboreality and the heavy forest exploitation within their range (Rijksen and Meijaard 1999). The Sumatran subpopulations appear to be more effectively connected through male dispersal for two reasons. First, the uninhabited mountain regions connecting subpopulations are forested, and thus dispersing males, who have been sighted at altitudes of up to 2000 m above sea level (Rijksen 1978), can move through them. Second, Sumatran forests provide suitable habitat to higher altitudes than Bornean ones due to the *Massenerhebung* effect (van Schaik et al. 1995), and this makes it easier for migrating males to cross rivers at their headwaters.

The example of the Sumatran orangutan demonstrates that even species with a geographically very limited range can show strong underlying genetic structure, caused by geographical barriers, habitat discontinuities, limited dispersal, and long population persistence. Correspondingly, genetic diversity might be mainly found among local subpopulations rather than within, and local extinctions carry a serious risk of losing a substantial part of a species' total genetic diversity. Our study highlights the need to assess the genetic make-up of endangered species in detail, identify local subpopulation boundaries, and focus conservation efforts on maintaining dispersal corridors among genetic clusters.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Funding

Swiss National Science Foundation (3100A-116848 to M.K. and C.vS.); Messerli Foundation; A.H.-Schultz Foundation; Claraz Schenkung.

Acknowledgments

We are very grateful to Ellen Meulman, Andrea Permana, Izumi, Zulfikar, Tony Weingrill, Helga Peters, Gregoire Bertagnolio, Gail Campbell-Smith, Yenny Saraswati and Rachmad Wahyudi for collecting samples. Corinne Ackermann and Kai Ansmann performed valuable laboratory work for this study. We thank Erik Willems for contributing the map of northern Sumatra. Furthermore, we highly appreciate the support of the following institutions: Indonesian State Ministry of Research and Technology (RISTEK), Indonesian Institute of Sciences (LIPI), Sabah Wildlife Department, Taman National Gunung Leuser (TNGL), Borneo Orangutan Survival Foundation (BOSF), Leuser International Foundation (LIF), and Badan Pengelola Kawasan Ekosistem Leuser (BPKEL).

References

Alter SE, Rynes E, Palumbi SR. 2007. DNA evidence for historic population size and past ecosystem impacts of gray whales. Proceedings of the National Academy of Sciences of the United States of America 104:15162–15167.

Anthony NM, Johnson-Bawe M, Jeffery K, et al. 2007. The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. Proceedings of the National Academy of Sciences of the United States of America 104:20432–20436.

Arora N, Nater A, van Schaik CP, et al. 2010. Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (*Pongo pygmaeus*). Proceedings of the National Academy of Sciences 107:21376–21381.

Arora N, van Noordwijk MA, van Schaik CP, et al. 2012. Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in the non-gregarious Asian great apes, the Bornean orangutans *(Pongo pygmaeus)* Mol Ecol. 21:3352–3362.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol and Evol. 16:37–48.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of chimpanzee populations. PLoS Genet. 3:e66. doi:10.1371/journal.pgen.0030066.

Bengtsson BO. 1978. Avoiding inbreeding—at what cost? Journal of Theor Biol. 73:439–444.

Bergl RA, Vigilant L. 2007. Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (*Gorilla gorilla diehli*). Mol Ecol. 16:501–516.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10:195–205.

Chesner CA, Rose WI, Deino A, Drake R, Westgate JA. 1991. Eruptive history of Earth's largest Quaternary caldera (Toba, Indonesia) clarified. Geology. 19:200–203.

Delgado RA, Van Schaik CP. 2000. The behavioral ecology and conservation of the orangutan (*Pongo pygmaeus*): a tale of two islands. Evol Anthrop. 9:201–218.

Dobson FS. 1982. Competition for mates and predominant juvenile male dispersal in mammals. An Behav. 30:1183–1192.

Douadi MI, Gatti S, Levrero F, et al. 2007. Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). Mol Ecol. 16:2247–2259.

Eriksson J, Hohmann G, Boesch C, Vigilant L. 2004. Rivers influence the population genetic structure of bonobos (*Pan* paniscus). Mol Ecol. 13:3425–3435.

Eriksson J, Siedel H, Lukas D, et al. 2006. Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (*Pan paniscus*). Mol Ecol. 15:939–949.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 14: 2611–2620.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 10:564–567.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes—application to human mitochondrial DNA restriction data. Genetics. 131:479–491.

Galdikas BMF. 1995. Social and reproductive behavior of wild adolescent female orangutans. In: Nadler RD, Galdikas BFM, Sheeran LK, Rosen N, editors. The neglected ape. New York: Plenum Press. p. 163–182.

Gaveau DLA, Wich S, Epting J, et al. 2009. The future of forests and orangutans (*Pongo abelii*) in Sumatra: predicting impacts of oil palm plantations, road construction, and mechanisms for reducing carbon emissions from deforestation. Environmental Research Letters 4. doi:10.1088/1748-9326/4/3/034013

Gonder MK, Locatelli S, Ghobrial L, et al. 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. Proceedings of the National Academy of Sciences of the United States of America 108:4766–4771.

Goossens B, Chikhi L, Jalil MF, et al. 2005. Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. Mol Ecol. 14:441–456.

Groves CP. 2001. Primate taxonomy. Washington, DC/London: Smithsonian Institution Press.

Guschanski K, Caillaud D, Robbins MM, Vigilant L. 2008. Females shape the genetic structure of a gorilla population. Current Biol. 18:1809–1814.

Handley LJL, Perrin N. 2007. Advances in our understanding of mammalian sex-biased dispersal. Mol Ecol. 16:1559–1578.

Hedrick PW, Kalinowski ST. 2000. Inbreeding depression in conservation biology. Annu Rev Ecol Syst. 31:139–162.

Husson SJ, Wich SA, Marshall AJ, et al. 2009. Orangutan distribution, density, abundance and impacts of disturbance. In: Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP, editors. Orangutans: geographic variation in behavioral ecology and conservation. Oxford: Oxford University Press.

IUCN. 2012. IUCN Red List of Threatened Species. Version 2012.1. http://www.iucnredlist.org, last accessed August 24, 2012.

Jalil MF, Cable J, Inyor JS, et al. 2008. Riverine effects on mitochondrial structure of Bornean orang-utans (*Pongo pygmaeus*) at two spatial scales. Mol Ecol. 17:2898–2909.

Johnson ML, Gaines MS. 1990. Evolution of dispersal—theoretical models and empirical tests using birds and mammals. Annu Rev Ecol Syst. 21:449–480.

Kanthaswamy S, Kurushima JD, Smith DG. 2006. Inferring *Pongo* conservation units: a perspective based on microsatellite and mitochondrial DNA analyses. Primates. 47:310–321.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. Ecology Letters 7:1225–1241.

Kimura M, Weiss GH. 1964. Stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

Knott CD. 1998. Changes in orangutan caloric intake, energy balance, and ketones in response to fluctuating fruit availability. Intl J Primatol. 19:1061–1079.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics. 22:768–770.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20:86–93.

Mitani JC, Watts DP, Muller MN. 2002. Recent developments in the study of wild chimpanzee behavior. Evol Anthrop. 11:9–25.

Morin PA, Chambers KE, Boesch C, Vigilant L. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). Mol Ecol. 10:1835–1844.

Morrogh-Bernard HC, Morf NV, Chivers DJ, Krutzen M. 2011. Dispersal patterns of orang-utans (*Pongo* spp.) in a Bornean peat-swamp forest. Intl J Primatol. 32:362–376.

Muir CC, Galdikas BMF, Beckenbach AT. 2000. mtDNA sequence diversity of orangutans from the islands of Borneo and Sumatra. Journal of Mol Evol. 51:471–480.

Nater A, Nietlisbach P, Arora N, et al. 2011. Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant orangutans (genus: *Pongo*). Mol Biol Evol. 28:2275–2288.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nietlisbach P, Arora N, Nater A, et al. 2012. Heavily male-biased long-distance dispersal of orang-utans (*genus: Pongo*), as revealed by Y-chromosomal and mitochondrial genetic markers. Mol Ecol. 21:3173–3186.

Nietlisbach P, Nater A, Greminger MP, Arora N, Krützen M. 2010. A multiplex-system to target 16 male-specific and 15 autosomal genetic markers for orang-utans (*genus: Pongo*). Conserv Genet Resour. 2:153–158.

Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res. 22:201–204.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol Biol and Evol. 25:1253–1256.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Pusey A, Wolf M. 1996. Inbreeding avoidance in animals. Trends Ecol Evol. 11:201–206.

Pusey AE. 1987. Sex-biased dispersal and inbreeding avoidance in birds and mammals. Trends Ecol Evol. 2:295–299.

Reed DH, Frankham R. 2003. Correlation between fitness and genetic diversity. Conser Biol. 17:230–237.

Rijksen HD. 1978. A field study on Sumatran orangutans (*Pongo pygmaeus abelii* Lesson 1827): ecology, behavior and conservation. Wageningen (The Netherlands): H. Veenman and Zonen.

Rijksen HD, Meijaard E. 1999. Our vanishing relative: the status of wild orang-utans at the close of the twentieth century. Dordrecht (The Netherlands): Kluwer Academic Publishers.

Roman J, Palumbi SR. 2003. Whales before whaling in the North Atlantic. Science 301:508–510.

Ross KG. 2001. Molecular ecology of social behaviour: analyses of breeding systems and genetic structure. Mol Ecol. 10:265–284.

Schlötterer C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma. 109:365–371.

Setia TM, van Schaik CP. 2007. The response of adult orang-utans to flanged male long calls: inferences about their function. Folia Primatologica. 78:215–226.

Singleton I, van Schaik CP. 2002. The social organisation of a population of Sumatran orang-utans. Folia Primatologica. 73:1–20.

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. Science. 236:787–792.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics. 139:457–462.

Soares P, Ermini L, Thomson N, et al. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Human Genet. 84:740–759.

Steiper ME. 2006. Population history, biogeography, and taxonomy of orangutans (Genus: *Pongo*) based on a population genetic meta-analysis of multiple loci. J Human Evol. 50:509–522.

Storz JF. 1999. Genetic consequences of mammalian social structure. J Mammal. 80:553–569.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512–526.

Tata HL, van Noordwijk M, Mulyoutami E, et al. 2010. Human livelihoods, ecosystem services and the habitat of the Sumatran orangutan: rapid assessment in Batang Toru and Tripa. Bogor (Indonesia): World Agroforestry Centre (ICRAF) Southeast Asia Regional Office.

Tautz D, Gerloff U, Hartung B, Fruth B, Hohmann G. 1999. Intracommunity relationships, dispersal pattern and paternity success in a wild living community of Bonobos (*Pan paniscus*) determined from DNA analysis of faecal samples. Proceedings of the Royal Society of London Series B—Biological Sciences. 266:1189–1195.

Utami Atmoko SS, Singleton I, van Noordwijk MA, van Schaik C, Setia TM. 2009. Male-male relationships in orangutans. In: Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP, editors. Orangutans: geographic variation in behavioral ecology and conservation. Oxford: Oxford University Press.

van Noordwijk MA, Arora N, Willems EP, et al. 2012. Female philopatry and its social benefits among Bornean orangutans. Behav Ecol and Sociobiol. 66:823–834.

van Schaik CP, Mirmanto E. 1985. Spatial variation in the structure and litterfall of a Sumatran rain forest. Biotropica. 17:196–205.

van Schaik CP, Monk KA, Robertson JMY. 2001. Dramatic decline in orangutan numbers in the Leuser ecosystem, Northern Sumatra. Oryx. 35:14–25.

van Schaik CP, Priatna A, Priatna D. 1995. Population estimates and habitat preferences of orangutans based on line transects of nests. In: Nadler RD, Galdikas BFM, Sheeran LK, Rosen N, editors. The Neglected Ape. New York: Plenum Press. p. 129–147.

Verstappen HT. 1973. A geomorphological reconnaissance of Sumatra and adjacent islands (Indonesia). Groningen (The Netherlands): Wolters-Noordhoff B.V.

Warren KS, Verschoor EJ, Langenhuijzen S, et al. 2001. Speciation and intrasubspecific variation of Bornean orangutans, *Pongo pygmaeus pygmaeus*. Mol Biol and Evol. 18:472–480.

Wich S, Riswan, Jenson J, Refish J, Nelleman C. 2011. Orangutans and the economics of sustainable forest management in Sumatra. UNEP/GRASP/PanEco/YEL/ICRAF/GRID-Arendal. http://www.unep.org/pdf/orangutan_report_scr.pdf, last accessed August 24, 2012.

Wich SA, Buij R, van Schaik C. 2004. Determinants of orangutan density in the dryland forests of the Leuser ecosystem. Primates. 45:177–182.

Wich SA, de Vries H, Ancrenaz M, et al. 2009. Orangutan life history variation. In: Wich SA, Utami Atmoko SS, Mitra Setia T, van Schaik CP, editors. Orangutans: geographic variation in behavioral ecology and conservation. Oxford: Oxford University Press.

Wich SA, Meijaard E, Marshall AJ, et al. 2008. Distribution and conservation status of the orang-utan (*Pongo* spp.) on Borneo and Sumatra: how many remain? Oryx. 42:329–339.

Williams GC. 1966. Adaptation and natural selection: a critique of some current evolutionary thought. Princeton (NJ): Princeton University Press.

Wilson GA, Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics. 163:1177–1191.

Wright S. 1943. Isolation by distance. Genetics. 28:114–138.

Received February 3, 2012; Revised July 9, 2012; Accepted July 10, 2012

Corresponding Editor: Adalgisa Caccone

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 447–460 doi: 10.1111/1755-0998.12075

Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny

B. NUSSBERGER,* M. P. GREMINGER,† C. GROSSEN,* L. F. KELLER* and P. WANDELER*
*Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, Zurich, CH 8057, Switzerland, †Anthropological Institute & Museum, University of Zurich, Winterthurerstrasse 190, Zurich, CH 8057, Switzerland

Abstract

Introgression can be an important evolutionary force but it can also lead to species extinction and as such is a crucial issue for species conservation. However, introgression is difficult to detect, morphologically as well as genetically. Hybridization with domestic cats (*Felis silvestris catus*) is a major concern for the conservation of European wildcats (*Felis s. silvestris*). The available morphologic and genetic markers for the two *Felis* subspecies are not sufficient to reliably detect hybrids beyond first generation. Here we present a single nucleotide polymorphism (SNP) based approach that allows the identification of introgressed individuals. Using high-throughput sequencing of reduced representation libraries we developed a diagnostic marker set containing 48 SNPs (F_{ST} > 0.8) which allows the identification of wildcats, domestic cats, their hybrids and backcrosses. This allows assessing introgression rate in natural wildcat populations and is key for a better understanding of hybridization processes.

Keywords: conservation genetics, *Felis silvestris*, genetic marker, hybridization, introgression, reduced representation library

Received 9 July 2012; revision received 4 December 2012; accepted 11 December 2012

Introduction

Introgression is difficult to detect, yet it is an important issue in evolutionary biology and conservation. Introgression, the flow of genes between taxa through hybridization beyond the first generation of hybrids (F1), can be an important evolutionary force (Grant *et al.* 2004; Seehausen 2004; Grant & Grant 2009) and can also lead to species extinction (Rhymer & Simberloff 1996). Introgression is especially a conservation concern when it is anthropogenic (Allendorf *et al.* 2001). This is the case in the crosses between European wildcats (*Felis silvestris silvestris*) and domestic cats (*Felis s. catus*). To assess the threat caused by hybridization, we need to quantify the introgression rate in potentially threatened populations. Therefore, it is crucial to overcome the difficulties in detecting not only F1, but also introgressed individuals, which are the decisive hybrids from a conservation perspective (Allendorf *et al.* 2001).

Introgression is difficult to detect for several reasons. First, morphological criteria are frequently not useful, since hybrids beyond the F1 generation are morphologically often indistinguishable from the parental species (Barbour *et al.* 2007; Krüger *et al.* 2009; Seiler *et al.* 2009; Ostberg *et al.* 2011). In wildcats for example, even the distinction between parentals of both hybridizing taxa based on morphology alone has been questioned (Daniels *et al.* 1998; Nussberger & Weber 2007). Second, the genetic identification is challenging because introgressed individuals share a large part of their genome with one of the parental species. For instance, a first generation backcross shares on average 75% of its genes with the parental species in which it has backcrossed. Consequently, many genetic markers are required to detect the presence of genes from the other parental species, especially when markers are highly polymorphic and not diagnostic, e.g. microsatellites (Vähä & Primmer 2006). For backcross detection, single nucleotide polymorphism (SNP) markers appear promising. SNPs are mostly biallelic (Lai 2001) and they cannot be more than tetrallelic (A, C, G or T). Due to the low number of alleles and the low degree of homoplasy, SNP markers are more likely to be diagnostic than highly polymorphic markers. Therefore, SNP markers are particularly useful for detecting introgressed hybrids. For example, diagnostic SNP markers have been used to detect introgression in hybridizing fish taxa (Finger *et al.* 2009; Simmons *et al.* 2009; Hohenlohe *et al.* 2011; Amish *et al.* 2012).

Here we report a SNP-based approach that allows the identification of introgressed individuals, and we illustrate it with data from European wildcats. European

Correspondence: Beatrice Nussberger, Fax: +41 (0)44 635 68 18; E-mail: Beatrice.nussberger@ieu.uzh.ch

wildcats are known to hybridize and to have fertile offspring with domestic cats (Beaumont *et al.* 2001; Randi *et al.* 2001; Pierpaoli *et al.* 2003; Lecis *et al.* 2006; Oliveira *et al.* 2008b; Hertwig *et al.* 2009; O'Brien *et al.* 2009). Hybridization with domestic cats is considered one of the major threats to the wildcat in many European countries (Driscoll & Nowell 2010). There is a need to better recognize and understand the processes and extent of introgression in order to develop appropriate conservation measures. However, the microsatellite marker sets commonly used to distinguish between wildcats and domestic cats have limited power to distinguish introgressed individuals, that is, hybrids beyond the F1 generation (Oliveira *et al.* 2008a, b; Hertwig *et al.* 2009; Say *et al.* 2012). In Hertwig *et al.* (2009), 3.5% of simulated F2 and 47% of simulated backcrosses were misinterpreted as parentals. In Oliveira *et al.* (2008b), 12% of simulated F2 and 20% of simulated backcrosses were erroneously attributed to parentals. Clearly, a set of more powerful markers is needed to assess the level of introgression in natural wildcat populations and the degree of threat to wildcats. Here, we aimed to obtain a set of diagnostic SNP markers for identifying wildcats, domestic cats, as well as their hybrids and backcrosses, by identifying single nucleotide polymorphisms in the genome where wild- and domestic cats present markedly different allele frequencies, using high-throughput sequencing of reduced representation libraries and Sanger sequencing.

Materials and methods

Methodological strategy

In a first step we defined what we considered domestic cats and wildcats (*reference samples*), using morphology and genetic data of a total of 45 potential domestic cats and 33 potential wildcats. Subsequently we sequenced a small part of the genome (*reduced representation library*) of six wildcats and three domestic cats. The comparison between the sequences of wildcats and domestic cats revealed SNPs between both subspecies. We then selected 200 SNPs at which our wildcat and domestic cat samples showed differently fixed alleles (*SNP selection*). To validate their diagnostic value, these SNPs were genotyped in an additional ten wildcats and 13 domestic cats by Sanger sequencing (*SNP validation*). Finally, we tested if our markers can assess the hybrid status of simulated individuals with known hybrid status (*SNP power assessment*).

Reference samples

Domestic cat samples (blood, gonads, hairs) were provided by Swiss veterinary offices and private cat owners ($n = 35$). We assumed that all these cats were domestic, since they lived in close proximity with humans and were tame. Moreover, most of the domestic cat samples came from regions where wildcats are absent. Eleven of these domestic cats were purebred. In addition to these domestic cats, gamekeepers provided samples from ten stray cats with domestic phenotype, from regions where wildcats occur (Supporting Information Table S1).

Blood or tissue samples from potential wildcats of the Swiss Jura region were provided by the Centre for Fish and Wildlife Health in Berne, Switzerland, by gamekeepers and by the Natural History Museums of Basel, Berne, La Chaux-de-Fonds, Lausanne, Neuchatel and Olten ($n = 33$, Table S1). We defined the reference wildcats according to both genetic and morphologic criteria. We followed the genetic identification method suggested by Driscoll *et al.* (2011), with a modified set of markers. We genotyped all potential wildcats at 24 autosomal microsatellites (Menotti-Raymond *et al.* 1999) and one Y linked microsatellite (Luo *et al.* 2007). In addition, we sequenced 2698 bp (base pair) of the mitochondrial DNA genes ND5 and ND6 (Driscoll *et al.* 2007) and 376 bp of SRY and 366 bp of SMCY on the Y chromosome (Pecon-Slattery *et al.* 2004; King *et al.* 2007). A complete list of markers with their primer sequences is provided in Table S1. For comparison, we further generated the same genetic data for 30 domestic cats from various breeds. Population substructure among all wildcats and these 30 domestic cats was identified using STRUCTURE (Pritchard *et al.* 2000). Mitochondrial DNA and Y-chromosomal haplotypes sequence data were compared with published haplotype sequences (Driscoll *et al.* 2007) in GENEIOUS PRO 4.8.5 Software (Drummond *et al.* 2009) to ascertain wildcat specificity. We only considered samples as reference wildcats if the animals carried wildcat mtDNA and Y haplotypes and if the proportion of autosomal wildcat ancestry (q value) was ≥ 0.95 according to STRUCTURE. Furthermore, when pictures from the sampled wildcats were available, we checked if the genetic results corresponded to the classic morphologic criteria: permanent dorsal line stopping at base of tail, blunt tail tip, distinct tail bands, four stripes on nape, two stripes on shoulder, blurry broken stripes on flanks, rhinarium with upper black margin and gularis with white areola (Ragni & Possenti 1996; Kitchener *et al.* 2005).

Reduced representation library RRL

To achieve high enough coverage for SNP detection with a given amount of sequencing effort, we chose to sequence only a small portion (2%) of the genome. To this end we constructed reduced representation libraries (RRL) by digesting genomic DNA and size selecting fragments (Van Tassell *et al.* 2008). Genomic DNA was

extracted from six reference wildcat samples and three domestic cat samples (Biosprint 96 DNA Blood kit; Qiagen). The six wildcats used for implementing the RRL were selected to have different geographical origins throughout the Swiss Jura region, thus reducing the likelihood of having related individuals in the sample. To construct RRLs, we digested 25 μg of genomic DNA with 250 Units of HaeIII (New England Biolabs). We separated the digested genomic DNA on a Spreadex EL1200 Wide Mini S-2×4 gel (Elchrom Scientific) in a SEA 2000 electrophoresis chamber at 55 °C, 120 Volt, in 1×TAE running buffer (Elchrom Scientific), during 3 h. We excised fragments between 587 bp and 622 bp. To extract DNA fragments, we placed gel pieces in a dialysis membrane (Carl Roth, 1785.1 Dialysierschlauch Visking, Cellulose) filled with 1×TAE buffer and closed it with plastic clips (Carl Roth, H277.1 ZelluTrans/Roth Verschlussklammer). The membrane packages were placed in an electrophoresis chamber (SEA 2000) at 55 °C, 120 Volt, in 1×TAE running buffer, during approximately 45 min. We purified the eluate using the MinElute PCR Purification Kit (Qiagen) according to the manufacturer's protocol. We prepared the sequencing library and individually barcoded our samples following the instructions of the SOLiD™ 4 System Library Preparation Guide (Applied Biosystems, 2010). The sequencing library was only amplified with eight PCR cycles to minimize overamplification. After DNA quantification with qPCR (SOLiD™ 4 System Library Quantitation with the SOLiD™ Library TaqMan® Quantitation Kit; Applied Biosystems), each sample was diluted to 500 pM. We submitted pooled libraries to the Functional Genomics Center Zurich (FGCZ) who performed paired-end (50/35) sequencing on SOLiD 4 (Applied Biosystems).

SNP selection

Raw sequence reads from SOLiD 4 platform were mapped to the cat genome assembly version Felcat4 (Pontius *et al.* 2007) using the default settings in BIOSCOPE version 1.3.1 (Life Technologies). SNPs were called using DIBAYES (Life Technologies) with high and medium stringency settings. To be able to compare genotypes of all individuals at a given SNP site, SAMTOOLS version 0.1.12a (Li *et al.* 2009) was used in cases where no call was made by DIBAYES to check whether the SNP site was not sequenced or homozygous for the reference allele.

From these SNPs, we selected potentially diagnostic SNPs based on three criteria. First, SNPs had to be sequenced to at least ten times coverage in all samples. Second, SNPs had to be fixed for a different allele in wildcats and domestic cats, meaning that the polymorphism at the SNP was only found between and not within subspecies. Third, we only selected markers which were on different chromosomes or at least 10 kb from one another, since unlinked markers are best for hybrid detection. We verified fixed SNPs visually with the Integrative Genomics Viewer (Robinson *et al.* 2011).

SNP validation

As we only obtained genomic data of nine cats in our initial SNP detection, we verified the allelic state of 200 potentially diagnostic SNPs in up to 23 additional cats by PCR and Sanger sequencing, thereby generating a total of 32 reference cat genotypes (16 wildcats and 16 domestic cats). For each locus, we therefore designed PCR primer pairs (PRIMER 3, Rozen & Skaletsky 2000) to obtain PCR products of 200–799 bp encompassing these potentially diagnostic SNPs (Table S3). PCR conditions were 30 cycles with annealing at 59 °C (57 °C for SNP082). We sequenced these products using Big Dye Terminate v3.1 chemistry on a 3730 DNA Analyzer (Applied Biosystems). Subsequently we analysed sequence data with Sequencing Analysis 5.1. (Applied Biosystems) and edited them in GENEIOUS. The number of individuals to be sequenced per locus was determined by calculating the F_{ST}-values between wild- and domestic cats with the individuals already analysed. When F_{ST} was <0.7 after sequencing eight or 16 individuals, we did not further analyse this locus. F_{ST}-values were calculated as the difference between the expected heterozygosity in wild- and domestic cats taken together and the mean of the expected heterozygosity in wild- and domestic cats separately, divided by the expected heterozygosity in wild- and domestic cats taken together (Conner & Hartl 2004).

SNP power assessment

We wanted to assess the power of the 48 SNP markers with highest F_{ST}-values (>0.8) in determining the correct hybrid status of simulated hybrids. To simulate hybrid genotypes, we needed genotypes for parental wildcats and domestic cats. To identify parental cats we first genotyped these 48 SNP markers in 42 additional cats, which had not been used to classify the markers based on F_{ST}-values. Using new genotypes avoids the 'high-grading' bias in assessing power of marker sets described by Anderson (2010). The 42 additional individuals comprised 18 domestic cats, ten stray cats, seven reference wildcats and seven potential wildcats with unclear status due to a contradiction between mtDNA or Y marker and autosomal microsatellites. We then used the program NEWHYBRIDS VERSION 1.1 BETA (Anderson & Thompson 2002) to assess the posterior probability of

belonging to the following six categories for each of these 42 samples: parental wildcats (W), parental domestic cats (D), first generation hybrids (F1), second generation hybrids (F2, i.e. F1×F1), backcrosses into wildcats (B×W, i.e. F1 × W), backcrosses into domestic cats (B×D, i.e. F1 × D). We used the default parameters of the program NEWHYBRIDS and did not include any other individuals in this analysis than these 42 samples. All samples which had ≥ 0.95 posterior probability of belonging to the parental categories D or W were used as parental samples to simulate F1, F2 and backcrossed hybrids.

We created the genotypes of hybrids F1, F2 and backcrosses (B×D and B×W) by sampling without replacement from amongst the alleles in the parental samples, using R 2.9.2 (RDevelopementCoreTeam 2009). Sampling the parental alleles without replacement avoids the problem of simulating lots of hybrid individuals that all carry a copy of the same allele in the parental sample. However, it limits the number of hybrids that can be generated. We simulated as many hybrids and backcrosses as we had parental alleles to distribute. For example, with nine parental wildcats, we had 18 alleles to create 18 F1 (in combination with nine domestic cats, resp. 18 domestic alleles) or 12 B×W (in combination with six domestic alleles needed for six F1). We analysed the simulated hybrids in NEWHYBRIDS, each hybrid category separately. In each NEWHYBRIDS run, we included the genotypes of the defined pure 16 wildcats and 16 domestic cats that were used in the RRL and SNP validation steps as known parentals, using the z and s option of NEWHYBRIDS. We repeated the simulation and analysis steps 200 times for each hybrid category. We defined individuals as correctly assigned by NEWHYBRIDS when their true category was the category with the highest scaled likelihood. Scaled likelihoods are the posterior probabilities to belong to a certain hybrid category, under a model where a priori every one of the hybrid categories is equally likely. We calculated the percentage of correctly assigned individuals (accuracy) and the mean scaled likelihoods (posterior probabilities) of all simulated individuals per category.

Furthermore, to explore the extent to which hybridization beyond the second hybrid generation is detectable with our method, we simulated using four additional categories of hybrids: crosses between a backcross into wildcat and a parental wildcat (B×W × W) and between backcross into wildcat and F1 (B×W × F1) and the same for domestic cats (B×D × D, B×D × F1). We analysed simulated individuals of all ten categories separately with NEWHYBRIDS, allowing for ten genotype frequency classes (2 parentals, F1, F2, B×W, B×D, B×W × W, B×W × F1, B×D × D, B×D × F1).

Results

Reference samples

We identified 24 reference wildcats based on microsatellites, mtDNA and Y markers. We had pictures of 19 of these cats. The phenotype of all these cats fulfilled the usual wildcat criteria. Nine potential wildcats were possibly of admixed ancestry and thus were not considered as reference wildcats: two potential wildcats (WK050, WK054) showed evidence of possible introgression at the autosomal microsatellites ($q < 0.95$ in STRUCTURE), and seven potential wildcats were of wildcat ancestry at the nuclear markers with $q \geq 0.95$ but mtDNA or Y markers were of the domestic cat type. All 43 domestic cats and stray cats which were analysed with microsatellite markers, mtDNA and/or Y markers were confirmed as domestic cats (Table S1).

RRL

The sequencing of the reduced representation libraries of six wildcats and three domestic cats yielded 597 139 577 sequenced beads. About 48% of these beads (285 234 154), representing a total of 11.5 gigabases, could be mapped to the reference Cat Genome.

SNP selection

At 654 out of 876 690 called SNP positions, all RRL samples were sequenced at least ten times and were fixed for alternate alleles in domestic and wild cats. However, when these fixed SNPs were verified within the Integrative Genomics Viewer, several of these SNP positions contained additional alleles, although at low coverage and mostly with a low Phred quality score (<20). As a consequence we selected by eye the 200 SNPs displaying the lowest number of reads with an alternate allele (Table S3).

SNP validation

Table 1 shows the positions of 187 potentially diagnostic SNPs on the domestic cat reference genome (FelCat4 December 2008, Pontius *et al.* 2007) and gives the corresponding allele frequencies for wildcats and domestic cats. Differences in allele frequencies are graphically shown in Fig. 1. We excluded 13 markers (6.5%) because of primer mismatch, indel-allele or multiple product amplification (e.g. primer binding region in a repeated element, Table S3). Overall, F_{ST}-values for the SNPs ranged from zero to one. Fifty SNPs had an F_{ST}-value of >0.8 between wildcats and domestic cats, including seven SNPs with $F_{ST} = 1$ (Table 1, F_{ST}).

Table 1 List of 187 SNP markers to detect introgression in domestic cats (D) and wildcats (W). Chromo_Position indicates the position of the SNP on the cat reference Genome, FelCat4 (Pontius et al. 2007, version Dec. 2008). F_{ST} is a measure of genetic differentiation between D and W. p and q are the two alleles found at the SNP position. *n*D and *n*W indicate the number of D and W successfully genotyped at the SNP position. p in D and q in W represent the frequencies of the alleles in the two subspecies. Nr gives the rank of the SNP after sorting by F_{ST}, SNP Nr is the identification number

Nr	SNP Nr	Chromo_position	F_{ST}	p	q	*n*D	p in D	q in D	*n*W	p in W	q in W
1	33	C1_133254300	1	T	G	16	1	0	16	0	1
2	101	B4_143164026	1	C	T	16	1	0	16	0	1
3	129	B4_96741303	1	G	A	16	1	0	16	0	1
4	138	C2_142773339	1	G	A	16	1	0	16	0	1
5	149	A3_157140228	1	A	C	16	1	0	16	0	1
6	158	B3_37642991	1	C	T	16	1	0	16	0	1
7	187	D3_49022779	1	C	G	16	1	0	16	0	1
8	12	A3_90799249	0.94	G	T	16	0.97	0.03	16	0	1
9	102	C2_142858667	0.94	C	T	16	0.97	0.03	16	0	1
10	105	D2_106505320	0.94	C	T	16	0.97	0.03	16	0	1
11	107	E2_51498305	0.94	C	A	16	0.97	0.03	16	0	1
12	115	A2_63544109	0.94	G	A	16	1	0	16	0.03	0.97
13	141	E1_47366937	0.94	G	A	16	1	0	16	0.03	0.97
14	155	B2_129152112	0.94	A	C	16	1	0	16	0.03	0.97
15	178	C1_189621758	0.94	G	A	16	0.97	0.03	16	0	1
16	194	E1_125241814	0.94	C	T	16	0.97	0.03	16	0	1
17	196	E2_50523470	0.94	T	A	16	1	0	16	0.03	0.97
18	198	E3_13634364	0.94	T	C	16	1	0	16	0.03	0.97
19	18	B1_58403280	0.88	C	A	16	0.94	0.06	16	0	1
20	32	C1_118678562	0.88	C	G	16	0.94	0.06	16	0	1
21	62	D2_88876341	0.88	G	T	16	0.94	0.06	16	0	1
22	93	B3_28741053	0.88	C	T	16	0.94	0.06	16	0	1
23	109	F2_65362892	0.88	G	A	16	0.94	0.06	16	0	1
24	133	C1_163375181	0.88	G	A	16	1	0	16	0.06	0.94
25	139	D4_75458793	0.88	T	C	16	0.94	0.06	16	0	1
26	148	A2_120724549	0.88	G	A	16	0.94	0.06	16	0	1
27	162	B3_99865718	0.88	G	A	16	0.94	0.06	16	0	1
28	192	D4_51926783	0.88	G	A	16	0.94	0.06	16	0	1
29	193	D4_52053226	0.88	C	T	16	0.94	0.06	16	0	1
30	184	D2_2202956	0.88	C	T	16	0.97	0.03	16	0.03	0.97
31	195	E2_33320051	0.88	A	G	16	0.97	0.03	16	0.03	0.97
32	14	B1_123418311	0.83	A	G	16	0.91	0.09	16	0	1
33	28	B4_143439104	0.83	G	A	16	0.91	0.09	16	0	1
34	41	D4_37998587	0.83	T	C	16	0.91	0.09	16	0	1
35	48	A3_51056949	0.83	C	T	16	0.91	0.09	16	0	1
36	57	D1_98155760	0.83	T	C	16	0.91	0.09	16	0	1
37	58	D1_126067118	0.83	G	A	16	0.91	0.09	16	0	1
38	60	D1_128802001	0.83	A	T	16	0.91	0.09	16	0	1
39	65	D3_76217054	0.83	G	T	16	0.91	0.09	16	0	1
40	88	F2_9296568	0.83	A	G	16	0.91	0.09	16	0	1
41	146	A1_214220499	0.83	C	T	16	0.91	0.09	16	0	1
42	176	C1_112821482	0.83	T	C	16	0.91	0.09	16	0	1
43	189	D3_73181465	0.83	C	T	16	0.91	0.09	16	0	1
44	190	D3_88773687	0.83	C	G	16	0.91	0.09	16	0	1
45	20	B2_132559340	0.82	A	G	16	0.94	0.06	16	0.03	0.97
46	26	B3_75494376	0.82	G	C	16	0.94	0.06	16	0.03	0.97
47	30	B4_45476816	0.82	A	G	16	0.94	0.06	16	0.03	0.97
48	159	B3_39998169	0.82	A	C	16	0.94	0.06	16	0.03	0.97
49	166	B3_147841323	0.82	G	A	16	0.94	0.06	16	0.03	0.97
50	50	C1_223335334	0.82	G	T	15	0.90	0.10	15	0	1
51	98	E1_47901546	0.78	G	A	16	0.88	0.13	15	0	1
52	1	A1_214461789	0.78	G	C	16	0.88	0.13	16	0	1

Table 1 (Continued)

Nr	SNP Nr	Chromo_position	F_{ST}	p	q	*n*D	p in D	q in D	*n*W	p in W	q in W
53	64	D3_70959423	0.78	A	G	16	0.88	0.13	16	0	1
54	126	B3_102961557	0.78	G	C	16	0.88	0.13	16	0	1
55	67	E2_64389936	0.77	G	T	16	0.91	0.09	16	0.03	0.97
56	152	B1_168327330	0.77	G	A	16	0.91	0.09	16	0.03	0.97
57	127	B3_132539085	0.77	C	T	16	0.94	0.06	16	0.06	0.94
58	106	D4_36844519	0.76	A	C	15	0.87	0.13	16	0	1
59	21	B2_38455848	0.76	C	T	15	0.90	0.10	16	0.03	0.97
60	66	E2_28834826	0.76	G	A	10	1	0	14	0.14	0.86
61	177	C1_177165193	0.74	C	G	14	0.86	0.14	16	0	1
62	7	A2_36537402	0.74	G	T	14	0.89	0.11	14	0.04	0.96
63	38	D2_16797246	0.74	A	G	14	0.89	0.11	14	0.04	0.96
64	114	A2_62528310	0.74	G	A	14	0.89	0.11	14	0.04	0.96
65	90	A1_80251090	0.73	G	A	14	0.89	0.11	13	0.04	0.96
66	136	C2_10551765	0.73	G	A	13	0.85	0.15	13	0	1
67	27	B4_106165338	0.73	C	T	16	0.84	0.16	16	0	1
68	143	F2_29878116	0.73	C	T	16	0.84	0.16	16	0	1
69	96	D4_61706901	0.71	C	T	15	0.83	0.17	16	0	1
70	153	B2_11210402	0.70	G	A	14	0.82	0.18	14	0	1
71	36	D1_9247995	0.69	C	G	14	0.82	0.18	15	0	1
72	151	B1_57974383	0.69	C	T	9	0.83	0.17	13	0	1
73	17	B1_24516687	0.69	G	A	6	0.83	0.17	9	0	1
74	95	B4_106085849	0.69	T	G	6	0.83	0.17	9	0	1
75	89	F2_29604098	0.69	C	T	16	0.81	0.19	15	0	1
76	173	B4_122774768	0.68	T	C	16	0.81	0.19	16	0	1
77	6	A2_22115264	0.68	A	C	9	0.83	0.17	14	0	1
78	84	D4_103411241	0.68	A	G	14	0.86	0.14	14	0.04	0.96
79	10	A3_150434747	0.68	A	G	6	0.83	0.17	10	0	1
80	19	B2_11748866	0.68	G	A	6	0.83	0.17	10	0	1
81	37	D2_15700028	0.68	T	C	6	0.83	0.17	10	0	1
82	45	F1_24323263	0.68	T	C	6	0.83	0.17	10	0	1
83	56	D1_72733259	0.68	A	C	6	0.83	0.17	10	0	1
84	69	F1_31149992	0.68	C	T	6	0.83	0.17	10	0	1
85	72	F2_64410099	0.68	A	G	6	0.83	0.17	10	0	1
86	76	B3_3763474	0.68	G	A	6	0.83	0.17	10	0	1
87	80	C2_134622594	0.68	G	A	6	0.83	0.17	10	0	1
88	83	D4_60140710	0.68	G	A	6	0.83	0.17	10	0	1
89	100	A2_154972126	0.68	T	C	6	0.83	0.17	10	0	1
90	113	A1_267376697	0.68	A	G	6	0.83	0.17	10	0	1
91	160	B3_67119952	0.68	T	C	6	0.83	0.17	10	0	1
92	163	B3_104962724	0.68	C	G	6	0.83	0.17	10	0	1
93	167	B4_2696116	0.68	C	T	6	0.83	0.17	10	0	1
94	175	C1_88089878	0.68	T	C	6	0.83	0.17	10	0	1
95	199	F2_4630496	0.68	A	G	6	0.83	0.17	10	0	1
96	200	F2_21635256	0.68	A	G	6	0.83	0.17	10	0	1
97	63	D2_111465892	0.67	A	G	13	0.81	0.19	14	0	1
98	181	D1_999750	0.67	C	T	13	0.81	0.19	14	0	1
99	51	C2_64959967	0.66	G	A	6	0.92	0.08	10	0.10	0.90
100	134	C1_188295633	0.66	G	T	6	0.92	0.08	10	0.10	0.90
101	174	C1_6047515	0.66	T	C	6	0.92	0.08	10	0.10	0.90
102	168	B4_2713634	0.65	G	A	10	0.85	0.15	14	0.04	0.96
103	170	B4_44289069	0.65	T	A	10	0.85	0.15	14	0.04	0.96
104	171	B4_44832398	0.65	C	A	10	0.85	0.15	14	0.04	0.96
105	164	B3_130995527	0.65	A	C	15	0.83	0.17	16	0.03	0.97
106	8	A2_6906598	0.65	C	G	14	0.79	0.21	14	0	1
107	86	F1_85116491	0.64	T	C	6	0.83	0.17	13	0	1
108	15	B1_191096484	0.64	T	A	10	0.80	0.20	14	0	1
109	49	B4_147077847	0.64	C	T	12	0.83	0.17	16	0.03	0.97

Table 1 (Continued)

Nr	SNP Nr	Chromo_position	F_{ST}	p	q	*n*D	p in D	q in D	*n*W	p in W	q in W
110	82	D3_124203045	0.64	G	A	7	0.86	0.14	13	0.04	0.96
111	71	F2_45763245	0.63	G	T	8	0.81	0.19	14	0	1
112	111	A1_222959361	0.63	G	A	16	0.81	0.19	16	0.03	0.97
113	16	B1_20092839	0.62	A	G	10	0.90	0.10	14	0.11	0.89
114	74	A1_239785943	0.62	C	T	14	0.86	0.14	14	0.07	0.93
115	61	D1_129618021	0.61	C	T	10	0.95	0.05	14	0.18	0.82
116	52	C2_74163720	0.59	A	G	5	0.80	0.20	10	0	1
117	142	F1_20032493	0.59	G	A	5	0.80	0.20	10	0	1
118	197	E2_66138174	0.59	T	G	5	0.80	0.20	10	0	1
119	42	E1_72880071	0.59	G	A	6	0.83	0.17	10	0.05	0.95
120	188	D3_60909701	0.59	G	C	6	0.75	0.25	7	0	1
121	157	B3_6909289	0.58	C	G	12	0.83	0.17	14	0.07	0.93
122	122	B2_84861747	0.58	A	G	10	0.80	0.20	13	0.04	0.96
123	35	C2_45117916	0.57	C	T	10	0.80	0.20	14	0.04	0.96
124	180	C2_137811507	0.57	G	T	10	0.80	0.20	14	0.04	0.96
125	40	D3_32104510	0.57	T	C	16	0.81	0.19	15	0.07	0.93
126	54	C2_140733169	0.56	G	A	10	0.75	0.25	14	0	1
127	4	A1_9371605	0.54	A	G	6	0.75	0.25	10	0	1
128	11	A3_169913387	0.54	T	A	6	0.75	0.25	10	0	1
129	13	A3_93714149	0.54	T	C	6	0.75	0.25	10	0	1
130	22	B2_67536455	0.54	G	A	6	0.75	0.25	10	0	1
131	24	B3_57147258	0.54	A	G	6	0.75	0.25	10	0	1
132	55	D1_68082963	0.54	T	C	6	0.75	0.25	10	0	1
133	117	A3_28148083	0.54	C	T	6	0.75	0.25	10	0	1
134	147	A2_42383186	0.54	G	A	6	0.75	0.25	10	0	1
135	154	B2_71247052	0.54	T	C	6	0.75	0.25	10	0	1
136	165	B3_135866504	0.54	G	T	6	0.75	0.25	10	0	1
137	179	C2_68465481	0.54	G	A	6	0.75	0.25	10	0	1
138	191	D4_10426918	0.54	T	C	6	0.75	0.25	10	0	1
139	99	F1_26460636	0.53	C	T	4	0.75	0.25	7	0	1
140	183	D1_109313008	0.52	C	A	10	0.80	0.20	14	0.07	0.93
141	2	A1_269159716	0.51	G	A	6	0.83	0.17	10	0.10	0.90
142	29	B4_15403984	0.51	G	A	6	0.83	0.17	10	0.10	0.90
143	59	D1_128044982	0.51	C	G	6	0.83	0.17	10	0.10	0.90
144	156	B2_134892585	0.50	C	A	15	0.80	0.20	16	0.09	0.91
145	73	E3_33733408	0.50	G	A	4	0.75	0.25	8	0	1
146	47	F2_7927040	0.45	G	C	6	0.75	0.25	10	0.05	0.95
147	145	A1_151348480	0.44	C	A	4	0.75	0.25	10	0	1
148	104	C2_158469278	0.44	C	G	5	0.70	0.30	9	0	1
149	130	B4_111855682	0.41	A	G	5	0.70	0.30	10	0	1
150	132	C1_82808777	0.41	A	G	5	0.70	0.30	10	0	1
151	3	A1_274277184	0.41	A	G	6	0.67	0.33	10	0	1
152	46	F2_3749961	0.41	C	A	6	0.67	0.33	10	0	1
153	79	C1_30344863	0.41	A	G	6	0.67	0.33	10	0	1
154	81	D1_11065896	0.41	A	G	6	0.67	0.33	10	0	1
155	92	B1_202073444	0.41	A	G	6	0.67	0.33	10	0	1
156	103	C2_151794647	0.41	C	T	6	0.67	0.33	10	0	1
157	124	B3_77335049	0.41	A	G	6	0.67	0.33	10	0	1
158	128	B3_148360238	0.41	C	T	6	0.67	0.33	10	0	1
159	137	C2_11113978	0.41	G	A	6	0.67	0.33	10	0	1
160	182	D1_88915301	0.41	T	C	6	0.67	0.33	10	0	1
161	140	D4_104246955	0.38	C	T	6	0.75	0.25	10	0.10	0.90
162	116	A2_200475325	0.36	C	T	5	0.60	0.40	7	0	1
163	43	E2_23114722	0.34	A	G	5	0.70	0.30	10	0.05	0.95
164	44	E3_12301230	0.34	A	G	10	0.80	0.20	14	0.21	0.79
165	25	B3_73330050	0.33	T	C	6	0.67	0.33	10	0.05	0.95
166	131	C1_34406063	0.33	A	G	6	0.67	0.33	10	0.05	0.95

Table 1 (Continued)

Nr	SNP Nr	Chromo_position	F_{ST}	p	q	*n*D	p in D	q in D	*n*W	p in W	q in W
167	135	C1_207927310	0.33	G	A	6	0.67	0.33	10	0.05	0.95
168	87	F2_2358597	0.33	G	A	8	0.63	0.38	14	0	1
169	34	C1_50675581	0.33	C	A	6	0.75	0.25	10	0.15	0.85
170	112	A1_247553760	0.33	C	T	6	0.75	0.25	10	0.15	0.85
171	9	A3_143339672	0.31	G	T	3	0.67	0.33	7	0	1
172	5	A2_176836753	0.29	T	C	6	0.58	0.42	10	0	1
173	31	B4_80349376	0.29	A	C	6	0.58	0.42	10	0	1
174	75	A2_130163447	0.29	C	T	6	0.58	0.42	10	0	1
175	85	E3_12162520	0.29	T	A	6	0.58	0.42	10	0	1
176	91	A3_100831036	0.29	G	A	6	0.58	0.42	10	0	1
177	121	A3_126916218	0.29	C	T	6	0.58	0.42	10	0	1
178	161	B3_71735716	0.29	C	G	6	0.58	0.42	10	0	1
179	108	F2_18305725	0.23	A	C	6	0.58	0.42	10	0.05	0.95
180	68	E2_64946728	0.18	A	G	6	0.50	0.50	10	0	1
181	77	B3_140493835	0.18	G	C	6	0.50	0.50	10	0	1
182	185	D2_9756017	0.11	C	T	6	0.58	0.42	10	0.20	0.80
183	110	F2_68402465	0.08	A	G	5	0.60	0.40	10	0.25	0.75
184	78	B4_52463921	0.08	A	G	6	0.42	0.58	10	0	1
185	118	A3_31797110	0.08	A	T	6	0.42	0.58	10	0	1
186	119	A3_73505900	0.05	T	C	5	0.50	0.50	10	0.10	0.90
187	125	B3_78472523	0.05	T	G	6	0.42	0.58	10	0.05	0.95

SNP power assessment

Based on 48 nuclear SNP markers with F_{ST}-values >0.8, NEWHYBRIDS assigned 41 of the 42 additionally genotyped cats with >0.95 posterior probability to one of six possible categories. All 18 domestic cats and ten stray cats were classified as parental domestic cats. Three reference wildcats (WK026, WK041, WK045) and one wildcat with domestic Y marker (WK024) were classified as backcrosses into wildcat. Three reference wildcats (WK017, WK035, WK049) and six wildcats with domestic mtDNA marker (WK020, WK022, WK027, WK036, WK055, WK077) were classified as parental wildcats. One reference wildcat (WK145) was classified as parental wildcat, but with a posterior probability of only 0.77 and was therefore excluded for hybrid simulation. Thus we had 28 parental domestic cats and nine parental wildcats to simulate hybrid genotypes.

NEWHYBRIDS assigned 99.6% of simulated individuals to the correct hybrid category with >0.50 posterior probability when using the 48 SNPs with highest F_{ST}-values (Table 2). 97.3% of the simulated individuals were assigned with >0.95 posterior probability to their true category. The mean posterior probabilities to belong to the true category was >0.98 for all simulated categories (Table 3).

Using only 32 of the SNPs with highest F_{ST}-values slightly lowered the mean posterior probabilities of belonging to either hybrid category, but, overall, still 98.6% of all individuals were correctly categorized. With 24 markers, the accuracy was still 97.7% (data not shown).

In the NEWHYBRIDS analysis of third generation hybrids, still 86.5% of simulated individuals were correctly assigned and the posterior probabilities for the ten simulated categories were around 0.8 (Tables 2 and 3). Eight percent of the parental domestic cats and 18% of the parental wildcats were erroneously categorized as third generation hybrids. However, in all hybrid categories, less than 1% of all simulated hybrids were assigned to the parental groups. Thus, while not all parental are correctly identified as such, hybrids are recognized correctly with high probability, although not always assigned to the correct hybrid category.

Discussion

First and second generation hybrids are reliably recognized with our set of SNP markers. We were able to identify the hybrid category of 97.3% of all simulated individuals with a posterior probability of >0.95, using 48 markers with highest F_{ST}-values (F_{ST} > 0.8). Even when including third generation hybrids, our marker set still allowed the correct identification of 86.5% of the simulated individuals. Thus, our new approach to detect SNP markers does work well in the case of the wildcats, domestic cats and their hybrids. Our approach consisted in sequencing a similar fraction of the genome of reference animals from both parental taxa, selecting SNPs diagnostic in these reference animals and verifying these SNPs in additional individuals. This marker development protocol will also be useful to find diagnostic SNPs in other hybridizing species.

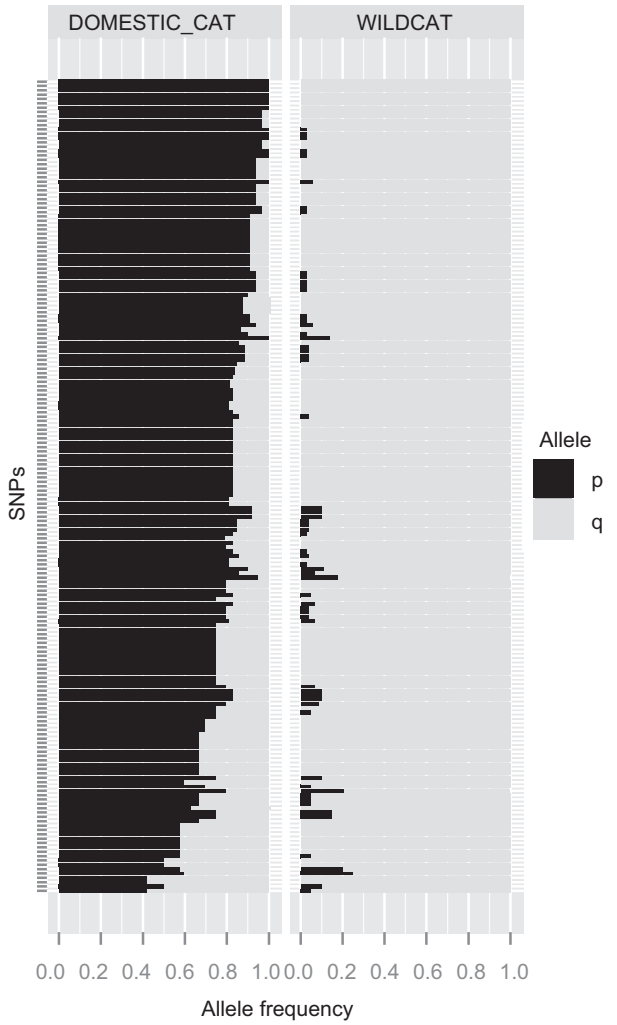


Fig. 1 Allele frequencies for both alleles p and q in domestic cats and wildcats at 187 SNP markers. Every horizontal bar represents one of the 187 SNP positions. The SNPs are ordered along the vertical axis according to decreasing F_{ST}-values between wildcats and domestic cats.

Choosing the right reference samples to develop diagnostic markers is crucial, yet challenging. First, reference samples should not contain any hybrids, as this will reduce the chances of correctly identifying diagnostic markers. Second, for the method to be broadly applicable, reference samples should be representative of the genetic diversity in the parental populations. Every wildcat found in Europe today is a potential hybrid, since domestic cats are thought to have spread in the area of the European Wildcat (*Felis silvestris silvestris*) since Roman times (Faure & Kitchener 2009). Ideally, we therefore would have developed the markers using wildcat samples from before Roman times, i.e. from more than 2500 years ago. But ancient DNA is of low quality and quantity (Hofreiter *et al.* 2001) and reduced representation libraries require DNA of high quality and quantity. Thus, we instead analysed modern samples with 24 autosomal markers, mtDNA sequences and Y markers. Samples without any sign of hybridization in all these markers were defined as reference wildcats. These reference wildcats formed a genetically distinct group relative to domestic cats. We minimized the probability of having introgressed individuals in our domestic cat reference sample by using mainly domestic cats from regions far from the habitat of wildcats in Switzerland (Jura region). To ensure adequate representation of genetic diversity in our reference samples we used domestic cats from different breeds and regions and we included wildcats from across their range in Switzerland. We cannot tell at present whether these markers are also applicable to wildcats beyond the Swiss borders. But we expect the differentiation between wildcats and domestic cats to be much higher than the differentiation between wildcat populations within Europe. Therefore, we hypothesize our marker set is also applicable to samples from outside Switzerland. Preliminary results of samples from France, Italy, Germany, Hungary and Bulgaria, which we genotyped with a 96×96 SNP genotyping chip (data not shown), support this hypothesis. Still, we would encourage researchers to test the markers in a larger set of known reference samples from other countries. Further, our markers are tested only for the subspecies *Felis s. silvestris* and *catus*. Their applicability to other *Felis s.* subspecies remains to be investigated.

Our simulations for the SNP power assessment are subject to potential bias. As some introgression between wildcats and domestic cats is expected, only clearly differentiated individuals were used as parental animals for the simulations of hybrid categories. As a consequence, the samples used for the simulations may be enriched with individuals more differentiated than average. This can lead to an overestimation of the SNP power for hybrid identification, because the detection of a hybrid is easier the more differentiated the two parental animals are. However, we expect this bias to be small here, given the strong differentiation of the SNPs between both subspecies.

High-throughput sequencing allows detecting a high number of markers at once and thus seems to be the method of choice for future marker development (Twyford & Ennos 2012). In addition, it is often sufficient to sequence only a small part of the genome (Davey *et al.* 2011), as we did here with RRL. Recently, a similar approach using RAD tags was described for SNP discovery in trouts (Hohenlohe *et al.* 2011). A slightly different approach of detecting diagnostic markers was chosen by Karlsson *et al.* (2011), who found genetic differences between farmed and wild Atlantic salmon based on a 7K SNP-chip. All these high-throughput sequencing approaches offer the advantage of generating markers that cover a broad range of the genome.

Table 2 Power assessment with NEWHYBRIDS using 48 SNP markers with highest F_{ST} values ($F_{ST} > 0.8$). Assignments to each hybrid category from a number n of simulated genotypes from the following categories: parental wildcat (W), parental domestic cat (D), F1, F1×F1 (F2), backcross into wildcat (B×W), backcross into domestic cat (B×D) and beyond second generation also B×D × D, B×D × F1, B×W × W and B×W × F1. Number of correct assignments are highlighted in bold. Accuracy gives the percentage of correct assignments. Each individual was assigned to the category for which the posterior probability was highest based on a NEWHYBRIDS analysis

		Category assessed with the highest probability											
	True category	D	W	F1	F2	B×D	B×W	B×D × D	B×W × W	B×D × F1	B×W × F1	*n*	Accuracy
Until second generation	D	**1793**	0	0	0	7	0	–	–	–	–	1800	99.6
	W	0	**1793**	0	0	0	7	–	–	–	–	1800	99.6
	F1	0	0	**3589**	10	1	0	–	–	–	–	3600	99.7
	F2	0	0	0	**3592**	5	3	–	–	–	–	3600	99.8
	B×D	0	0	0	12	**2388**	0	–	–	–	–	2400	99.5
	B×W	0	0	0	22	0	**2378**	–	–	–	–	2400	99.1
Until third generation (10 categories)	D	**1654**	0	0	0	0	0	146	0	0	0	1800	91.9
	W	0	**1474**	0	0	0	0	0	326	0	0	1800	81.9
	F1	0	0	**3588**	5	0	0	0	0	4	3	3600	99.7
	F2	0	0	18	**2664**	4	0	0	0	445	469	3600	74.0
	B×D	0	0	0	0	**2090**	0	107	0	203	0	2400	87.1
	B×W	0	0	1	0	0	**2068**	0	82	0	249	2400	86.2
	B×D × D	20	0	0	0	250	0	**2128**	0	2	0	2400	88.7
	B×W × W	2	0	0	0	0	282	0	**2110**	0	6	2400	87.9
	B×D × F1	0	0	0	304	104	0	0	0	**1992**	0	2400	83.0
	B×W × F1	0	0	0	300	0	62	0	0	2	**2036**	2400	84.8

Table 3 Mean posterior probabilities and 99% confidence intervals of belonging to a defined hybrid category for a number n of simulated genotypes. Categories are: parental wildcat (W), parental domestic cat (D), F1, F1×F1 (F2), backcross into wildcat (B×W), backcross into domestic cat (B×D) and for the simulations of hybrids beyond second generation B×D × D, B×D × F1, B×W × W and B×W × F1. Values for the correct categories are highlighted in bold.

		Mean posterior probabilities and 99% confidence intervals										
	True category	D	W	F1	F2	B×D	B×W	B×D × D	B×W × W	B×D × F1	B×W × F1	*n*
Until second generation (10 categories)	D	**0.994 ± 0.003**	0 ± 0	0 ± 0	0 ± 0	0.006 ± 0.003	0 ± 0	–	–	–	–	1800
	W	0 ± 0	**0.994 ± 0.003**	0 ± 0	0 ± 0	0 ± 0	0.006 ± 0.003	–	–	–	–	1800
	F1	0 ± 0	0 ± 0	**0.993 ± 0.002**	0.006 ± 0.002	0.001 ± 0.001	0 ± 0	–	–	–	–	3600
	F2	0 ± 0	0 ± 0	0 ± 0	**0.995 ± 0.002**	0.003 ± 0.002	0.002 ± 0.001	–	–	–	–	3600
	B×D	0 ± 0	0 ± 0	0 ± 0	0.012 ± 0.003	**0.988 ± 0.003**	0 ± 0	–	–	–	–	2400
	B×W	0 ± 0	0 ± 0	0 ± 0	0.016 ± 0.005	0 ± 0	**0.984 ± 0.005**	–	–	–	–	2400
Until third generation (10 categories)	D	**0.879 ± 0.013**	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.121 ± 0.013	0 ± 0	0 ± 0	0 ± 0	1800
	W	0 ± 0	**0.789 ± 0.019**	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.21 ± 0.019	0 ± 0	0 ± 0	1800
	F1	0 ± 0	0 ± 0	**0.991 ± 0.003**	0.004 ± 0.001	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.003 ± 0.001	0.002 ± 0.001	3600
	F2	0 ± 0	0 ± 0	0.005 ± 0.003	**0.642 ± 0.011**	0.001 ± 0.001	0 ± 0	0 ± 0	0 ± 0	0.171 ± 0.01	0.18 ± 0.01	3600
	B×D	0 ± 0	0 ± 0	0 ± 0	0.002 ± 0	**0.789 ± 0.013**	0 ± 0	0.074 ± 0.009	0 ± 0	0.136 ± 0.011	0 ± 0	2400
	B×W	0 ± 0	0 ± 0	0 ± 0.001	0.001 ± 0	0 ± 0	**0.798 ± 0.013**	0 ± 0	0.056 ± 0.008	0 ± 0	0.145 ± 0.012	2400
	B×D × D	0.017 ± 0.005	0 ± 0	0 ± 0	0 ± 0	0.147 ± 0.012	0 ± 0	**0.833 ± 0.013**	0 ± 0	0.003 ± 0.001	0 ± 0	2400
	B×W × W	0 ± 0	0.002 ± 0.002	0 ± 0	0 ± 0	0 ± 0	0.151 ± 0.013	0 ± 0	**0.843 ± 0.014**	0 ± 0	0.004 ± 0.002	2400
	B×D × F1	0 ± 0	0 ± 0	0 ± 0	0.191 ± 0.012	0.062 ± 0.009	0 ± 0	0.001 ± 0.001	0 ± 0	**0.742 ± 0.013**	0.004 ± 0.001	2400
	B×W × F1	0 ± 0	0 ± 0	0 ± 0	0.19 ± 0.012	0 ± 0	0.041 ± 0.007	0 ± 0	0 ± 0	0.005 ± 0.001	**0.765 ± 0.013**	2400

Different genetic questions need different genetic markers (Sunnucks 2000; Freeland 2005). Reliably recognizing hybrids beyond F1 has proven difficult with highly polymorphic microsatellite markers in several species (Fur seal: Kingston & Gwilliam 2007; Wildcats: Oliveira *et al.* 2008a; Hertwig *et al.* 2009; Say *et al.* 2012; Florida bog frogs: Austin *et al.* 2011). In theory as few as four to five fully diagnostic markers would be sufficient to identify recent backcrosses (Boecklen & Howard 1997). In our data, 24 almost diagnostic SNP markers were sufficient to correctly categorize 97.7% of all simulated hybrids, using a threshold for posterior probability of >0.5. However, with highly polymorphic, non diagnostic microsatellites, it takes about 48 markers to recognize backcrossed individuals with a posterior probability of >0.5 (Vähä & Primmer 2006). Most of the studies of hybridization in wildcats used between nine and 27 microsatellite markers, with allelic richness between seven and 43 (Beaumont *et al.* 2001; Randi *et al.* 2001; Pierpaoli *et al.* 2003; Lecis *et al.* 2006; Oliveira *et al.* 2008b; O'Brien *et al.* 2009). Markers with high allelic richness, like microsatellites, are well suited to recognize genetic population structure (Guichoux *et al.* 2011). However, high allelic richness in combination with homoplasy reduces the diagnostic power of markers for hybrid recognition, since there are more possibilities of allele sharing between two hybridizing taxa. Therefore, highly polymorphic markers developed for detecting genetic population structure are not the best markers to identify introgression. It is worth developing diagnostic markers with the explicit intent to detect introgression. The drawback of the diagnostic markers is that they should not be used for other genetic analyses such as genetic differentiation measures or paternity tests. On the other hand, the RRL approach we used for the diagnostic marker development generates enough high-throughput sequencing data to allow the development of other markers for other purposes as well.

SNPs are powerful markers to detect introgression. Their power resides in the highly differentiated allele frequencies between hybridizing taxa. Although high discriminatory power can also be reached with microsatellites (Burgarella *et al.* 2009), SNP markers present several advantages over microsatellite markers. SNPs are mostly biallelic. In our screening of 200 regions around a potentially diagnostic SNP, we found over 360 SNPs. Only two of them were triallelic (in sequence of SNP091 and SNP136) and none were tetrallelic. At biallelic SNPs, a diploid has only three options per locus: homozygous for either of the alleles, or heterozygous. This makes hybrid detection straightforward, at least in fixed SNPs. Heterozygosity at all SNP positions indicate a F1 hybrid and an individual having a proportion of 75% of the alleles from one parental is most probably a

backcross into that parental group. SNPs have also several technical advantages over microsatellites. Results obtained in different laboratories are compatible without a need to calibrate them. SNP genotyping assays are easier to multiplex than microsatellites, because they do not rely on the detection of fragment length. Finally, SNP genotyping assays can be designed to be very short, e.g. using PCR products shorter than 100 bp, because only a single base position has to be determined. This allows working with highly fragmented DNA and low DNA quantities, as is found in faeces, hair or ancient samples (Morin & McCarthy 2007).

In the near future, we aim to genotype non-invasively collected hair samples from free ranging wildcats to assess the introgression rate of domestic cats into different European wildcat populations. Depending on levels of introgression, management plans for species conservation can then be developed (Allendorf *et al*. 2001). Overall, our set of novel SNP markers allows the reliable assessment of introgression levels in natural populations and thus will help improve our understanding of the process of hybridization and introgression.

Acknowledgement

We thank Dominique Waldvogel, Glauco Camenisch, Nicole Ponta and Johanna Kinnunen for their help in the lab, and the FGCZ, University of Zürich (Rémy Bruggmann, Andrea Patrignani), LifeTechnologies (Gerrit Kuhn) and Elchrom Scientific (Danilo Tait, Marco Leu, Oliver Schicht) for technical support. We are grateful to the gamekeepers, the Centre for Fish and Wildlife Health FIWI, University of Berne (Marie-Pierre Ryser, Manuela Weber), the Vetsuisse Faculty of University of Zurich (Godelind Wolf, Iris Reichler) and the Natural History Museums Basel (Raffael Winkler), Berne (Peter Lueps), La Chaux-de-Fonds (Sunila Sen-Gupta), Lausanne (Olivier Glaizot), Neuchatel (Martin Zimmerli) and Olten (Peter Flückiger) for providing cat samples. We thank Eric Anderson for help with NewHybrids and three anonymous reviewers for their helpful comments. This work was funded by Lotterie + Sport-Toto-Fonds Solothurn, Zürcher Tierschutz, University Research Priority Program, Service des forêts, de la faune et de la nature du canton de Vaud, Service de la Faune et de la Pêche de l'État de Genève and Stiftung Naturschutz und Wild.

Conflict of interest

The authors declare no conflict of interest.

References

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology & Evolution*, **16**, 613–622.

Amish SJ, Hohenlohe PA, Painter S *et al.* (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources*, **12**, 653–660.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, **10**, 701–710.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Austin JD, Gorman TA, Bishop D, Moler P (2011) Genetic evidence of contemporary hybridization in one of North America's rarest anurans, the Florida bog frog. *Animal Conservation*, **14**, 553–561.

Barbour RC, Potts BM, Vaillancourt RE (2007) Gene flow between introduced and native Eucalyptus species: morphological analysis of Tri-species and backcross hybrids involving E-nitens. *Silvae Genetica*, **56**, 127–133.

Beaumont M, Barratt EM, Gottelli D *et al.* (2001) Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, **10**, 319–336.

Boecklen WJ, Howard DJ (1997) Genetic analysis of hybrid zones: numbers of markers and power of resolution. *Ecology*, **78**, 2611–2616.

Burgarella C, Lorenzo Z, Jabbour-Zahab R *et al.* (2009) Detection of hybrids in nature: application to oaks (Quercus suber and Q-ilex). *Heredity*, **102**, 442–452.

Conner JK, Hartl DL (2004) *A Primer of Ecological Genetics*, 1st edn. Sinauer Associates, Inc., Massachusetts. pp. 57–58.

Daniels MJ, Balharry D, Hirst D, Kitchener AC, Aspinall RJ (1998) Morphological and pelage characteristics of wild living cats in Scotland: implications for defining the 'wildcat'. *Journal of Zoology*, **244**, 231–247.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Driscoll C, Nowell K (2010) Felis silvestris. In: IUCN 2011. IUCN Red List of Threatened Species. Version 2011.2. www.iucnredlist.org

Driscoll CA, Menotti-Raymond M, Roca AL *et al.* (2007) The near eastern origin of cat domestication. *Science*, **317**, 519–523.

Driscoll C, Yamaguchi N, O'Brien SJ, Macdonald DW (2011) A Suite of Genetic Markers Useful in Assessing Wildcat (Felis silvestris ssp.) - Domestic Cat (Felis silvestris catus) Admixture. *Journal of Heredity*, **102**, S87–S90.

Drummond AJ, Ashton B, Cheung M, Heled J *et al.* (2009) Geneious v4.7, Available from http://www.geneious.com.

Faure E, Kitchener AC (2009) An archaeological and historical review of the relationships between felids and people. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals*, **22**, 221–238.

Finger AJ, Stephens MR, Clipperton NW, May B (2009) Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trouts. *Molecular Ecology Resources*, **9**, 759–763.

Freeland JR (2005) *Molecular Ecology*. John Wiley & Sons Ltd, Chichester.

Grant PR, Grant BR (2009) The secondary contact phase of allopatric speciation in Darwin's finches. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 20141–20148.

Grant PR, Grant BR, Markert JA, Keller LF, Petren K (2004) Convergent evolution of Darwin's finches caused by introgressive hybridization and selection. *Evolution*, **58**, 1588–1599.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Hertwig ST, Schweizer M, Stepanow S *et al.* (2009) Regionally high rates of hybridization and introgression in German wildcat populations (Felis silvestris, Carnivora, Felidae). *Journal of Zoological Systematics and Evolutionary Research*, **47**, 283–297.

Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S (2001) Ancient DNA. *Nature Reviews Genetics*, **2**, 353–359.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Karlsson S, Moen T, Lien S, Glover KA, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, **11**, 247–253.

King V, Goodfellow PN, Wilkerson AJP *et al.* (2007) Evolution of the male-determining gene SRY within the cat family Felidae. *Genetics*, **175**, 1855–1867.

Kingston JJ, Gwilliam J (2007) Hybridization between two sympatrically breeding species of fur seal at Iles Crozet revealed by genetic analysis. *Conservation Genetics*, **8**, 1133–1145.

Kitchener AC, Yamaguchi N, Ward JM, Macdonald DW (2005) A diagnosis for the Scottish wildcat (Felis silvestris): a tool for conservation action for a critically-endangered felid. *Animal Conservation*, **8**, 223–237.

Krüger M, Hertwig ST, Jetschke G, Fischer MS (2009) Evaluation of anatomical characters and the question of hybridization with domestic cats in the wildcat population of Thuringia, Germany. *Journal of Zoological Systematics and Evolutionary Research*, **47**, 268–282.

Lai E (2001) Application of SNP technologies in medicine: lessons learned and future challenges. *Genome Research*, **11**, 927–929.

Lecis R, Pierpaoli M, Biro ZS *et al.* (2006) Bayesian analyses of admixture in wild and domestic cats (Felis silvestris) using linked microsatellite loci. *Molecular Ecology*, **15**, 119–131.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Luo SJ, Johnson WF, David VA *et al.* (2007) Development of Y chromosome intraspecific polymorphic markers in the Felidae. *Journal of Heredity*, **98**, 400–413.

Menotti-Raymond M, David VA, Lyons LA *et al.* (1999) A genetic linkage map of microsatellites in the domestic cat (Felis catus). *Genomics*, **57**, 9–23.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes*, **7**, 937–946.

Nussberger B, Weber D (2007) The reliability of pelage characters for the diagnosis of the European Wildcat (Felis silvestris silvestris). In: *Felid Biology and Conservation Conference*, WildCRU, 100. University of Oxford, Oxford.

O'Brien J, Devillard S, Say L *et al.* (2009) Preserving genetic integrity in a hybridising world: are European Wildcats (Felis silvestris silvestris) in eastern France distinct from sympatric feral domestic cats? *Biodiversity and Conservation*, **18**, 2351–2360.

Oliveira R, Godinho R, Randi E, Alves PC (2008a) Hybridization versus conservation: are domestic cats threatening the genetic integrity of wildcats (Felis silvestris silvestris) in Iberian Peninsula? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2953–2961.

Oliveira R, Godinho R, Randi E, Ferrand N, Alves PC (2008b) Molecular analysis of hybridisation between wild and domestic cats (Felis silvestris) in Portugal: implications for conservation. *Conservation Genetics*, **9**, 1–11.

Ostberg CO, Duda JJ, Graham JH *et al.* (2011) Growth, morphology, and developmental instability of rainbow trout, yellowstone cutthroat trout, and four hybrid generations. *Transactions of the American Fisheries Society*, **140**, 334–344.

Pecon-Slattery J, Pearks Wilkerson AJ, Murphy WJ, O'Brien SJ (2004) Phylogenetic assessment of introns and SINEs within the y chromosome using the cat family felidae as a species Tree. *Molecular Biology and Evolution*, **21**, 2299–2309.

Pierpaoli M, Biro ZS, Herrmann M *et al.* (2003) Genetic distinction of wildcat (Felis silvestris) populations in Europe, and hybridization with domestic cats in Hungary. *Molecular Ecology*, **12**, 2585–2598.

Pontius JU, Mullikin JC, Smith DR *et al.* (2007) Initial sequence and comparative analysis of the cat genome. *Genome Research*, **17**, 1675–1689.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Ragni B, Possenti M (1996) Variability of coat-colour and markings system in Felis silvestris. *Italian Journal of Zoology*, **63**, 285–292.

Randi E, Pierpaoli M, Beaumont M, Ragni B, Sforzi A (2001) Genetic identification of wild and domestic cats (Felis silvestris) and their hybrids using Bayesian clustering methods. *Molecular Biology and Evolution*, **18**, 1679–1693.

RDevelopementCoreTeam (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available from http://www.R-project.org.

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Robinson JT, Thorvaldsdottir H, Winckler W *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology*, **29**, 24–26.

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds Krawetz S & Misener S), pp. 365–386. Humana Press, Totowa, NJ. Available from http://fokker.wi.mit.edu/primer3/.

Say L, Devillard S, Léger F, Pontier D, Ruette S (2012) Distribution and spatial genetic structure of European wildcat in France. *Animal Conservation*, **15**, 18–27.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, **19**, 198–207.

Seiler SM, Gunnell K, Ptacek MB, Keeley ER (2009) Morphological patterns of hybridization between yellowstone cutthroat trout and introduced rainbow trout in the South Fork of the Snake River watershed, idaho and wyoming. *North American Journal of Fisheries Management*, **29**, 1529–1539.

Simmons RE, Lavretsky P, May B (2009) Introgressive hybridization of redband trout in the upper McCloud River watershed. *Transactions of the American Fisheries Society*, **139**, 201–213.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, **15**, 199–203.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–189.

Vähä JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.

Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

B.N., P.W., and L.F.K. designed research; B.N. performed research; M.P.G. contributed new analytical tools; C.G. ran the simulations.

Data Accessibility

The following data are available on Dryad doi:10.5061/dryad.270b7.

mtDNA sequences: Fasta file containing mtDNA sequences of wild- and domestic cats.

MicrosatelliteData: Excel file containing microsatellite fragment length data for all individuals.

MorphologyData: Excel file containing description of diagnostic morphology criteria for all individuals.

RefSamplesSequences: Folder containing the raw sequences (subfolder SNPsequences) and the annotated consensus sequences (subfolder SNPconsensusSequences) for each of the 200 diagnostic SNPs.

SNPgenotypingData: Excel file containing SNP genotyping results for all individuals.

SimulationFiles: Folder containing R script for hybrid simulations (SimsNewHybrids.R), input files (SelectionSimulOnlyRefs040912.txt, SelectionSimulOnlyTestInd040912.txt) and file that holds the definitions of ten of the genotype frequency classes possible after three generations

of mating between two species (TwoGensGtypFreq10.txt).

Y_data: Folder containing the raw sequences from SRY sequencing (SNPseqSRY) and the SMCY microsatellite fragment length data for male individuals (SMCYdata).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Individuals: description of individuals (origin, morphology, Y, mtDNA, microsatellites).

Table S2 Markers: list of primers used to define reference samples.

Table S3 SNPprimers: list of primers used for SNP genotyping by sequencing.

Conservation Genet Resour (2010) 2:153–158
DOI 10.1007/s12686-010-9278-2

TECHNICAL NOTE

A multiplex-system to target 16 male-specific and 15 autosomal genetic markers for orang-utans (genus: *Pongo*)

Pirmin Nietlisbach · Alexander Nater · Maja P. Greminger · Natasha Arora · Michael Krützen

Received: 30 June 2010 / Accepted: 6 July 2010 / Published online: 25 July 2010

Abstract Genetic studies of dispersal on local spatial and short temporal scales require a large number of autosomal microsatellites. However, the study of dispersal over large spatial scales and the resolution of deep evolutionary histories require marker systems that are preferentially inherited through the male or female line. Addressing such questions in endangered orang-utans (genus: *Pongo*) bears significant relevance to species conservation, as habitat destruction and fragmentation pose a significant threat to the whole genus. Here, we report 16 male-specific markers (nine human-derived microsatellites, six single nucleotide and one insertion-deletion polymorphisms), and 15 novel *Pongo*-derived autosomal microsatellite loci. All 31 markers can be amplified in four multiplex polymerase chain reactions even in DNA derived from faecal material. The markers can be applied to studying a wide range of important questions in this genus, such as conservation genetics, social structure, phylogeny and phylogeography.

Keywords *Pongo* spp. · Single nucleotide polymorphisms · Microsatellites · Y chromosome · SNP typing · Non-invasive samples

Electronic supplementary material The online version of this article (doi:10.1007/s12686-010-9278-2) contains supplementary material, which is available to authorized users.

P. Nietlisbach · A. Nater · M. P. Greminger · N. Arora · M. Krützen (✉)
Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
e-mail: michael.kruetzen@aim.uzh.ch

The endangered orang-utans occur on the islands of Borneo (*Pongo pygmaeus*; about 50,000 animals) and Sumatra (*P. abelii*; about 6,500 animals), where they have undergone a recent dramatic decline in population size (Goossens et al. 2006; Wich et al. 2008). This has been mostly attributed to habitat loss, leading to heavily fragmented populations of often only a few hundred individuals (Wich et al. 2008). Therefore, it is essential to maintain genetic diversity, which has been linked to population fitness (e.g. Reed and Frankham 2003). This can be achieved by maintaining corridors between fragmented populations, allowing animals to follow natural dispersal patterns (Gilbert-Norton et al. 2010).

Studying natural dispersal in wild orang-utans pose significant challenges. Behavioural observations suggested higher male than female dispersal (Delgado and van Schaik 2000), although this has not been fully confirmed by previous genetic studies (Utami et al. 2002; Goossens et al. 2005), where patterns of direct dispersal were investigated using autosomal microsatellite markers. However, direct inferences from autosomal markers are limited to the timescale of a few generations and geographically small areas, as sexual recombination will break down sex-specific information (Goudet et al. 2002). Sex-biased dispersal over larger time and spatial scales can be investigated by contrasting genetic information obtained from markers inherited through either the male or female lineage (Handley and Perrin 2007).

In orang-utans, maternally transmitted mitochondrial DNA markers are widely available (e.g. Warren et al. 2001), but markers on the male-specific region of the Y chromosome have not yet been applied. Here, we report 16 male-specific markers for the application in the genus *Pongo*. Nine of these markers are human-derived microsatellite loci, six are single nucleotide polymorphisms (SNPs) and one is an

insertion-deletion polymorphism (indel). We also provide 15 novel orang-utan specific autosomal microsatellite markers, which yielded more reliable amplification results for use with highly degraded faecal samples than human-derived markers (data not shown).

In a previous study (Erler et al. 2004), 45 male-specific human-derived loci were shown to amplify in orang-utans. Of those, we selected 25 loci re-sequenced them, designed orang-utan specific primers, and tested for levels of polymorphism in 14 Bornean and Sumatran orang-utans (Table 1; see laboratory procedures below). After aligning flanking regions of all 25 microsatellite loci, we discovered two SNPs and one indel. Additionally, we identified four more SNPs by sequencing Y chromosome conserved anchored tagged sequence loci DBY13, SMCY12, and SMCY14 (Hellborg and Ellegren 2003). We combined 16 Y-specific polymorphic markers (GenBank Acc.No. HM803995-HM804006) into two multiplex PCRs (Table 1). Actual repeat number was linked to amplicon length through sequence data (Table S1 in the Online Resources). For SNP typing, we applied a modified fragment length discrepant allele specific PCR (Li et al. 2009). Each PCR combined a fluorescently labelled forward primer with two interrogating reverse primers. The reverse primers were complementary to one of each SNP allele at their 3′-end and differed in their length at the 5′-end; the resulting difference in product length allowed for allelic discrimination by fragment length analysis (Fig. 1a). One microsatellite locus had a SNP in close proximity and another a SNP and an indel, which allowed for simultaneous screening in one or two PCRs, respectively (Fig. 1). Multiplex reactions for Y markers consisted of 1 µl DNA template in an 8 µl volume, and were carried out with the Qiagen PCR Multiplex Kit according to manufacturer's instructions and varying primer concentrations (Table 1). In addition to the manufacturer's recommendations, we also included a touchdown step in the PCRs, starting at 4°C above the reported annealing temperature in Table 1, then decreasing by 0.5°C per cycle for eight cycles to reach the annealing temperature for the last 40 cycles. We assessed the variability of the Y-markers on a population level in 21 and 14 orang-utans from one Bornean (subspecies *P. pygmaeus wurmbii*) and one Sumatran site, respectively. In addition, we report the number of Y-polymorphisms in an extensive sample of 173 animals covering most of the range of the genus *Pongo* (Table 1, Nietlisbach 2009). We tested scoring reliability by repeating 32% of all genotyping reactions, and found an average scoring error rate of 0.6% over all loci. It has not escaped our notice that our results immediately suggest a possible diagnostic system for elucidating species affiliation or hybrid status for males of this genus

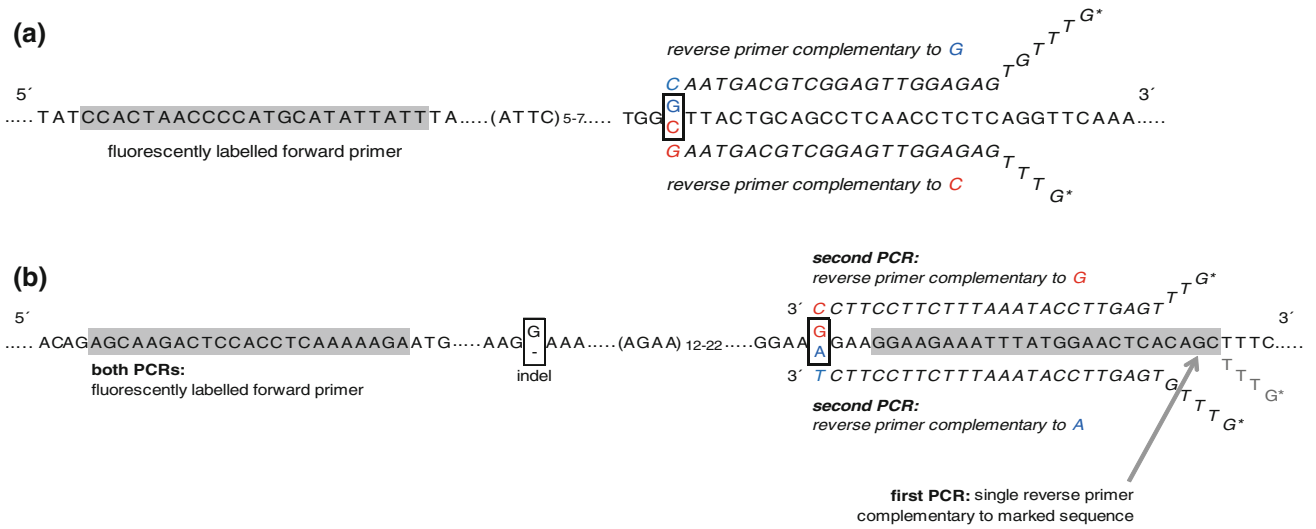


Fig. 1 Schematic representation of genotyping techniques for combined microsatellite/SNP typing. **a** Genotyping of SNP and microsatellite locus (DYS577). This locus was typed by a modified fragment length discrepant allele specific PCR (see text, Li et al. 2009) which targets the SNP in the box and the repeat motif of the microsatellite in one PCR. Discrimination between the two states of the SNP was achieved through competition of two reverse primers differing in length; and both reverse primers having a 1 to 2°C lower melting temperature and higher concentration than the fluorescently labelled forward primer (Table 1) to increase competition. Calibration was done with samples of known sequence. **b** Typing of three markers in two PCRs (DYS630). A first PCR using the grey primers amplified the region containing all three markers. A second PCR discriminates between SNP states as described in Fig. 1a. Due to confounding effects of the polymorphic microsatellite and indel, amplicon sizes of the first and second reactions have to be compared. If the amplicon of the first reaction is 5 bases longer than that of the second reaction, the SNP allele at this locus is G; if the difference is 3 bases, the SNP state is A. * PIG-tails were added to the 5′-end of reverse primers to enhance adenylation (Brownstein et al. 1996)

Table 1 Primers for male-specific and autosomal markers in orang-utans

Locus	Primer sequence 5′–3′	PC	T_A	Polymorphism	Overall		Suaq (Sumatra)				Tuanan (Borneo)			
					N	N_A	N	N_A	H_O	H_E	N	N_A	H_O	H_E
DYS630	*PET*-AGCAAGACTCCACCTCAAAAAGA*	0.15	63	AGAA	172	11	14	3		0.66	21	4		0.69
	gtttGCTGTGAGTTCCATAAATTTCTTCC	0.20		indel	172	2	14	1		0	21	1		0
DYS587	*6FAM*-AAAATTACCTTCTTTGGAAAGTAGTATT	0.30	63	ATACA	166	8	14	1		0	19	1		0
	gtttGTTATTTCTGAGCAGGGTTTCTAAG	0.40												
DYS532	*NED*-AGCAGGATTCCCTCTAAAAATATCA	0.10	63	compound,	171	3	14	1		0	21	2		0.09
	gTTTCTCCCTCCCTCCCTCTC	0.14		main motif (CTTT)										
DYS577	*6FAM*-CCACTAACCCCATGCATATTATT	0.30	63	GAAT	171	2	14	1		0	19	1		0
	gtttGAGAGGTTGAGGCTGCAGTAAG	0.40		C/G	171	2	14	2		0.13	19	1		0
	gtttgtGAGAGGTTGAGGCTG CAGTAAC	0.40												
DYS645	*6FAM*-GTACTAATTTTATTTCTTATGGCGTAGA	0.15	63	GTTTT	173	2	14	1		0	21	1		0
	gtttACACATGGCACCTGACACTG	0.20												
Y6C2	*6FAM*-CTTCTCCTTCTTCCTATTCTTCATCT	0.10	63	TTC	172	2	14	1		0	20	1		0
	gtttCAATAGTTTGGGGAAATAAGACAATG	0.14												
DBY13	*6FAM*-GGAAACTAAAATATGACATTGTAATTTG	0.30	63	C/G	168	2	14	1		0	20	1		0
	gtttATTTTTTTATTGTGATGCATACAGC	0.40												
	gtttgtATTTTTTTATTGTGATGCATACAGG	0.40												
DYS510	*PET*-GAAAGATAGATCAACAAGGTAGAAACAA	0.30	64	GATA	169	6	12	4		0.6	21	2		0.44
	gtttCATCCATCCATCCATCCATCT	0.40												
DYS561	*6FAM*-CCTGATGCCATCTGAAAATTAA	0.30	64	TAGA	168	5	14	1		0	20	3		0.52
	gtttACAACTGCACTCCAGCTTAGG	0.40												
DYS556	*6FAM*-TTACAAAACTAACATAAAGACCAAACAG	0.30	64	TAAA	172	3	14	1		0	21	2		0.41
	gtttGAAGCATTGAGTATAGTATAAAGTTGGT	0.40												
DYS630	*PET*-AGCAAGACTCCACCTCAAAAAGA*	0.15	64	A/G	171	2	14	1		0	21	1		0
	gttTGAGTTCCATAAATTTCTTCCTTCC	0.20												
	gtttgTGAGTTCCATAAATTTCTTCCTTCT	0.20												
SMCY12_26	*6FAM*-AAGGGTCACCACAGAAATACTTAG	0.15	64	C/G	173	2	14	2		0.13	21	1		0
	gtttgtCAGGTGGGGCGTAGTCTC	0.20												
	gtttCAGGTGGGGCGTAGTCTG	0.20												
SMCY12_337	*6FAM*-GTTACAGGTATACATGCACCTTTTT	0.15	64	A/C	171	2	14	1		0	21	1		0
	gtttgtGTGTGGCTTCTTTACTCTGTCA	0.20												
	gtttGTGTGGCTTCTTTACTCTGTCC	0.20												
SMCY14	*6FAM*-ATGGGGAAAAGATGAGTTCTGA	0.15	64	C/T	173	2	14	1		0	21	2		0.41
	gtttGTCTGGCATCCTAATGCCT	0.20												
	gtttgtGTCTGGCATCCTAATGCCC	0.20												

Table 1 continued

Locus	Primer sequence 5′–3′	PC	T_A	Polymorphism	Overall		Suaq (Sumatra)				Tuanan (Borneo)			
					N	N_A	N	N_A	H_O	H_E	N	N_A	H_O	H_E
O4_6	*PET*-GGCAATGTAACATATCCCTCTGTGT	0.05	58	GATA			23	4	0.61	0.68	28	3	0.71	0.62
	AGCCATGGACCTTGTGAGAAAAG	0.05												
O4_A5	*6FAM*-ATGGGCCCAGAAAACAACTCAGT	0.15	58	(GATA)(GATG)			22	4	0.64	0.55	26	6	0.65	0.65
	AGATAAAGGAATGGATAGATGGACAGA	0.15												
O4_A7	*VIC*-ACTGGCCCATTCAAAGTCTGTCATT	0.10	58	GTAG			21	4	0.86	0.72	26	2	0.35	0.29
	ACTGGCCCATTCAAAGTCTGT	0.10												
O4_A8	*NED*-CACAGGGTCCAAACTCAGATTATTG	0.20	58	(GATA)(GATG)			23	3	0.30	0.31	29	1	0	0
	CCTTCCCCCTCATGTAGTTATCAA	0.20												
O4_B5	*VIC*-GAGCCCTGATTCGTTTTACTGG	0.20	58	GATA			22	6	0.86	0.73	28	5	0.50	0.54
	AGCAAAGGCAGAAAACTGTAATGA	0.20												
O4_B6	*6FAM*-TGGAGCCTGAATATGTGACTGAAT	0.20	58	(GATA)(GTAG)			20	6	0.65	0.61	26	6	0.46	0.79
	AATGCCAGGATTTCCTTCTTTTT	0.20												
O4_B24	*6FAM*-TCTGAGGTACCCTGTAACAAAGAAA	0.10	58	GATA			23	3	0.65	0.56	29	1	0.00	0.00
	GAAATCCCAGTACCATATAAATGTCAT	0.10												
O4_A1	*6FAM*-CTCCCCTTCCTTCCTTTATTCAGTT	0.10	62	GTAG			23	5	0.87	0.73	28	4	0.82	0.75
	CAACACTTGGCAGTCACAAATCAG	0.10												
O4_B3	*VIC*-TTCCAGAAGGGGCGAGAAGTT	0.10	62	GACA			22	3	0.59	0.64	26	1	0	0
	GTTGGGACCAAACAGTTGTCAATAA	0.10												
O4_B17	*PET*-GTACCGACGGTGCACGAACAATGTA	0.30	62	GATG			19	3	0.68	0.67	26	6	0.69	0.73
	AGCCTGGCTGAAAAGTGGAACTGAG	0.30												
O4_B20	*NED*-CCTGCATTTTGTCACTCCCTCAACC	0.20	62	GATG			14	1	0	0	24	2	0.33	0.38
	CTGCCACACCTCCATGGACACAGAT	0.20												
O4_C9	*6FAM*-TGCAGGCCAGGGCTTCTTTCAA	0.15	62	GATA			22	5	0.55	0.54	27	4	0.59	0.63
	CAGTCTCCCCAGGACCCCTACACAG	0.15												
O4_C13	*6FAM*-CTGGGCACACTGTATATGGGGTAG	0.20	62	GATA			20	3	0.75	0.56	21	4	0.38	0.59
	GTTTGAGACCACTCATGATGCAAAGACC	0.20												
O4_Chr5	*PET*-CAGCAGCTCCTGAAATATCTGTCC	0.15	62	GATA			21	4	0.81	0.70	23	5	0.74	0.74
	GTTTGGGGTAGAGGAAAGCAGGTTGAT	0.15												
O4_Chr7	*NED*-CATCTCTTTATGGCTGACTGTTGAT	0.10	62	GATA			17	11	0.76	0.83	24	15	0.92	0.91
	GTTTGGTCCAAGACAAATTTGTATGAGT	0.10												

All loci with a Y in the name are Y-linked, all others are autosomal. Summary statistics are given for two study sites and over all sampled orang-utans. For loci DYS630 and DYS577, three and two male-specific markers were typed, respectively (Fig. 1). Loci combined in a single multiplex reaction have the same annealing temperature

Fluorescent labels are shown in italics at the 5′ end of the forward primer. PIG-tail bases (Brownstein et al. 1996) are given in lower case

PC primer concentration [μM], T_A annealing temperature, *N* number of samples, N_A number of different alleles, H_O observed heterozygosity, H_E expected heterozygosity (Nei 1987). * primer used in two PCRs (Fig. 1b)

Amplicon sizes and their relation to repeat numbers are shown in Table S1 in the Online Resources

(Nietlisbach 2009), if used in unison with readily available mtDNA markers.

To clone autosomal microsatellite markers, we extracted genomic DNA from 25 mg of frozen muscle tissue from a Sumatran orang-utan, using the DNeasy Tissue Kit (Qiagen). We digested ten micrograms of the purified DNA with *Nhe*I and *Alu*I (New England Biolabs) and size-selected for fragments between 400 and 1,200 base pairs length. Enrichment, cloning and sequencing were carried out as described in Nater et al. (2008), using only tetranucleotide biotinylated probes [$(GACA)_7$, $(GATA)_7$, and $(GATC)_7$]. We sequenced plasmids from 68 positive clones, of which 70% contained a microsatellite repeat. For 25 loci, which contained long uninterrupted repeats, we designed primers and amplified these loci in twelve orang-utans. Levels of polymorphism were qualitatively assessed on high-resolution Spreadex gels (Elchrom Scientific). Based on these results, we fluorescently labelled the forward primers of the 15 most polymorphic markers and combined these 15 loci (GenBank Acc.No. HM804007-HM804021) into two multiplex PCRs (Table 1). Then, we genotyped 29 orang-utans from Borneo and 23 from Sumatra, using DNA extracts from faecal samples with target DNA concentration ranging from 25 to 1,000 pg/µl, strictly following guidelines from Morin et al. (2001). PCRs using the Qiagen PCR Multiplex Kit contained 1 µl template DNA in an 8 µl final volume, with varying primer concentrations and annealing temperatures (Table 1). PCRs included 45 cycles with conditions according to manufacturer's instructions.

If not indicated otherwise, we used standard laboratory techniques at each step. We designed PCR primers with the PrimerSelect software implemented in Lasergene v7 (DNASTAR). PCR amplifications were performed on Veriti 96-well thermal cyclers (Applied Biosystems). Sequencing reactions were carried out using the BigDye Terminator v3.1 on a 3730 DNA Analyzer (both Applied Biosystems) according to manufacturer's instructions, cleaned-up using a $MgSO_4$ precipitation procedure, followed by resuspending the pellet in 20 µl ddH_2O. For fragment length analysis, PCR products were diluted 20–80 times in ddH_2O. One microlitre of this was added to 9.93 µl HiDi formamide and 0.07 µl of GeneScan 500 LIZ Size Standard (both Applied Biosystems) and denatured for three minutes at 95°C. We ran the samples on a 3730 DNA Analyzer and obtained genotypes using GeneMapper software v4.0 (Applied Biosystems). For the statistical analyses, we used MStools v3.1 add-into Microsoft Excel (Park 2001) and Genepop v4.0 (Rousset 2008).

Fragment length discrepant allele specific PCR used as SNP typing technique proved to be a reliable and cost-efficient strategy to assess SNP variation. The possibility to combine this technique with conventional microsatellite fragment length analysis makes it a suitable method to include a small number of SNPs to complement an extensive microsatellite analysis. The polymorphic male-specific markers for orang-utans described here promise to be highly useful for population genetic and phylogenetic studies addressing questions about dispersal strategies, phylogeographic patterns, and comparisons with other molecular markers. The autosomal markers can be applied to investigate local dispersal or assess relatedness and paternity. Knowledge about such processes, in particular about natural dispersal strategies, is important for species conservation.

Acknowledgments We thank C. van Schaik, M. van Noordwijk, J. Pamungkas, and D. Perwitasari-Farajallah. We are also indebted to all individuals who helped collecting samples in the field. This study was funded by the Swiss National Science Foundation (31003A-116848 to MK), Messerli Foundation, A.H.-Schultz Stiftung, and Claraz Schenkung. We thank the Indonesian Institute of Sciences (LIPI), the Indonesian State Ministry for Research and Technology (RISTEK), and the Sabah Wildlife Department for granting permission to undertake this research. All sampling and transportation of samples was conducted in accordance with Indonesian, Malaysian and international regulations (CITES).

References

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. Biotechniques 20(6): 1004–1010

Delgado RA, van Schaik CP (2000) The behavioral ecology and conservation of the orang-utan (*Pongo pygmaeus*): a tale of two islands. Evol Anthropol 9(5):201–218

Erler A, Stoneking M, Kayser M (2004) Development of Y-chromosomal microsatellite markers for nonhuman primates. Mol Ecol 13(10):2921–2930. doi:10.1111/j.1365-294X.2004.02304.x

Gilbert-Norton L, Wilson R, Stevens JR, Beard KH (2010) A meta-analytic review of corridor effectiveness. Conserv Biol 24(3): 660–668. doi:10.1111/j.1523-1739.2010.01450.x

Goossens B, Chikhi L, Jalil MF, Ancrenaz M, Lackman-Ancrenaz I, Mohamed M, Andau P, Bruford MW (2005) Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. Mol Ecol 14(2):441–456. doi:10.1111/j.1365-294X.02421.x

Goossens B, Chikhi L, Ancrenaz M, Lackman-Ancrenaz I, Andau P, Bruford MW (2006) Genetic signature of anthropogenic population collapse in orang-utans. PLoS Biol 4(2):285–291. doi: e2510.1371/journal.pbio.0040025

Goudet J, Perrin N, Waser P (2002) Tests for sex-biased dispersal using bi-parentally inherited genetic markers. Mol Ecol 11(6):1103–1114

Handley LJL, Perrin N (2007) Advances in our understanding of mammalian sex-biased dispersal. Mol Ecol 16(8):1559–1578. doi:10.1111/j.1365-294X.2006.03152.x

Hellborg L, Ellegren H (2003) Y chromosome conserved anchored tagged sequences (YCATS) for the analysis of mammalian male-specific DNA. Mol Ecol 12(1):283–291

Li SZ, Wan HR, Ji HY, Zhou KY, Yang G (2009) SNP discovery based on CATS and genotyping in the finless porpoise

(*Neophocaena phocaenoides*). Conserv Genet 10(6):2013–2019. doi:10.1007/s10592-009-9882-4

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). Mol Ecol 10(7):1835–1844

Nater A, Krützen M, Lindholm AK (2008) Development of polymorphic microsatellite markers for the livebearing fish *Poecilia parac*. Mol Ecol Resour 8(4):857–860. doi:10.1111/j.1755-0998.2008.02090.x

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nietlisbach P (2009) Male-specific markers in orang-utans (*Pongo* spp.)—Dispersal and phylogeny. MSc thesis. University of Zurich

Park SDE (2001) Trypanotolerance in West African cattle and the population genetic effects of selection. University of Dublin, Dublin, Ireland

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. Conserv Biol 17(1):230–237

Rousset F (2008) GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8(1):103–106. doi:10.1111/j.1471-8286.2007.01931.x

Utami SS, Goossens B, Bruford MW, de Ruiter JR, van Hooff J (2002) Male bimaturism and reproductive success in Sumatran orang-utans. Behav Ecol 13(5):643–652

Warren KS, Verschoor EJ, Langenhuijzen S, Heriyanto, Swan RA, Vigilant L, Heeney JL (2001) Speciation and intrasubspecific variation of Bornean orang-utans, *Pongo pygmaeus pygmaeus*. Mol Biol Evol 18 (4):472-480

Wich SA, Meijaard E, Marshall AJ, Husson S, Ancrenaz M, Lacy RC, van Schaik CP, Sugardjito J, Simorangkir T, Traylor-Holzer K, Doughty M, Supriatna J, Dennis R, Gumal M, Knott CD, Singleton I (2008) Distribution and conservation status of the orang-utan (*Pongo* spp.) on Borneo and Sumatra: how many remain? Oryx 42(3):329–339. doi:10.1017/s003060530800197x

Curriculum Vitae

Name:	Maja Patricia MATTLE-GREMINGER
Date of birth:	18 April 1983
Citizenship:	Richterswil (ZH), Switzerland

Education

2009 – 2015	**Ph.D. in Evolutionary Biology**, University of Zurich Dissertation: *Unraveling the Evolutionary History of Orangutans (genus: Pongo)—The impact of Environmental Processes and the Genomic Basis of Adaptation*
2005 – 2007	**Master of Science in Biology**, University of Zurich Thesis: *The quest for the Y – Development and application of male-specific markers in orangutans and bottlenose dolphins*
2002 – 2005	**Bachelor of Science in Biology**, University of Zurich

Publications in international and peer-reviewed journals

12. **Greminger MP**, Nater A, Roos C, Goossens B, Gut M, Gut IG, van Schaik CP, Marques-Bonet T, and Krützen M. Power and necessity of incorporating male-specific genomic data in analyses of species with discordant sex-specific evolutionary histories: a case study in orangutans (genus: Pongo). *Systematic Biology* (in review).
11. Bilgin Sonay T, Carvalho T, Robinson M, **Greminger MP**, Krützen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Research*, doi:10.1101/gr.190868.115.
10. Nater A, **Greminger MP,** Arora N, van Schaik CP, Goossens B, Singleton I, Verschoor EJ, Warren KS, Krützen M (2015) Reconstructing the demographic history of orang-utans using approximate Bayesian computation. *Molecular Ecology* 24, 310-327.
9. **Greminger MP**, Stolting K, Nater A, Goossens B, Arora N, Bruggmann R, Patrignani A, Nussberger B, Sharma R, Kraus RH, Ambu L, Singleton I, Chikhi L, van Schaik C, Krützen M (2014) Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics* 15, 16.
8. Nussberger B, **Greminger MP**, Grossen C, Keller LF, Wandeler P (2013) Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources* 13, 447-460.
7. Nater A, Arora N, **Greminger MP**, van Schaik CP, Singleton I, Wich SA, Fredriksson G, Perwitasari-Farajallah D, Pamungkas J, Krützen M (2013) Marked population structure and recent migration in the critically endangered Sumatran orangutan (Pongo abelii). *Journal of Heredity* 104, 2-13.

6. Rotheray E, Lepais O, Nater A, Krützen M, **Greminger MP**, Goulson D, Bussiere L (2012) Genetic variation and population decline of an endangered hoverfly Blera fallax (Diptera: Syrphidae). *Conservation Genetics* 13, 1283-1291.

5. Arora N, Van Noordwijk MA, Ackermann C, Willems EP, Nater A, **Greminger MP**, Nietlisbach P, Dunkel LP, Utami Atmoko SS, Pamungkas J, Perwitasari-Farajallah D, Van Schaik CP, Krützen M (2012) Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (Pongo pygmaeus). *Molecular Ecology* 21, 3352-3362.

4. Rotheray EL, **Greminger MP**, Nater A, Krützen M, Goulson D, Bussière L (2012) Polymorphic microsatellite loci for the endangered pine hoverfly Blera fallax (Diptera: Syrphidae). *Conservation Genetics Resources* 4, 117-120.

3. Nietlisbach P, Nater A, **Greminger MP**, Arora N, Krützen M (2010) A multiplex-system to target 16 male-specific and 15 autosomal genetic markers for orang-utans (genus: Pongo). *Conservation Genetics Resources* 2, 153-158.

2. **Greminger MP**, Kruetzen M, Schelling C, Pienkowska–Schelling A, Wandeler P (2010) The quest for Y-chromosomal markers–methodological strategies for mammalian non-model organisms. *Molecular Ecology Resources* 10, 409-420.

1. **Greminger MP,** Schäfer MA, Nater A, Blanckenhorn WU, Krützen M (2009) Development of polymorphic microsatellite markers for the dung fly (Sepsis cynipsea). *Molecular Ecology Resources* 9, 1554-1556.